From Parity to Preference-based Notions of Fairness in Classification

Muhammad Bilal Zafar[1], Isabel Valera[2], Manuel Gomez Rodriguez[1], Krishna P. Gummadi[1], and Adrian Weller[2,3,4]

[1]Max Planck Institute for Software Systems (MPI-SWS), [2]University of Cambridge, [3]Alan Turing Institute, [4]Leverhulme Centre for the Future of Intelligence

Abstract

Many notions of fairness in data-driven decision making are inspired by the concept of discrimination in social sciences and law, and focus on ensuring *parity* (equality) in treatment or outcomes for different social groups. In this paper, we propose *preference*-based notions of fairness with the goals of avoiding potential 'reverse-discrimination' and enabling high decision accuracy. We introduce tractable proxies to design convex boundary-based classifiers that satisfy these new notions of fairness and show on the ProPublica COMPAS dataset that these notions allow for greater decision accuracy than parity-based fairness.

This paper is a shortened version of arxiv:1707.00010.

1 Introduction

As machine learning is increasingly being used to automate decision making in domains that affect human lives (*e.g.*, credit ratings, housing allocation, recidivism risk prediction), there are growing concerns about the potential for *unfairness* in such algorithmic decisions [16, 18]. A flurry of recent research on fair learning has focused on defining appropriate notions of fairness and then designing mechanisms to ensure fairness in automated decision making [8, 9, 12, 13, 14, 21, 22, 23].

Existing notions of fairness in the machine learning literature are largely inspired by the concept of **discrimination** in social sciences and law. These notions call for **parity** (*i.e.*, equality) in **treatment**, in **impact**, or both. To ensure parity in treatment (or treatment parity), decision making systems need to avoid using users' sensitive attribute information, *i.e.*, avoid using the membership information in socially salient groups (*e.g.*, gender, race), which are protected by anti-discrimination laws [2, 6]. As a result, the use of group-conditional decision making systems is often prohibited. To ensure parity in impact (or impact parity), decision making systems need to avoid disparity in the fraction of users belonging to different sensitive attribute groups (*e.g.*, men, women) that receive *beneficial* decision outcomes. A number of learning mechanisms have been proposed to achieve parity in treatment [8, 17], parity in impact [4, 12, 14] or both [9, 11, 13, 21, 22, 23]. However, these mechanisms pay a significant cost in terms of the accuracy (or utility) of their predictions. In fact, there exist some inherent tradeoffs (both theoretical and empirical) between achieving high prediction accuracy and satisfying treatment and/or impact parity [5, 7, 10, 15].

In this work, we introduce, formalize and evaluate new notions of fairness that are inspired by the concepts of **fair division** and **envy-freeness** in economics and game theory [3, 20]. Our work is motivated by the observation that, in certain decision making scenarios, the existing parity-based fairness notions may be too stringent, precluding more accurate decisions, which may also be desired by every sensitive attribute group. To relax these parity-based notions, we introduce the concept of a user **group's preference** for being assigned one set of decision outcomes over another. Given the choice between various sets of decision outcomes, any group of users would collectively *prefer* the set that contains *the largest fraction* (or the greatest number) of beneficial decision outcomes for that group.[1] More specifically, our new preference-based notions of fairness, which we formally define in the next section, use the concept of user group's preference as follows:

— **From Parity Treatment to Preferred Treatment:** To offer preferred treatment, a decision making system should ensure that every sensitive attribute group (*e.g.*, men and women) *prefers* the set of decisions they receive over the set of decisions they would have received had they collectively presented themselves to the system as members of a different sensitive group.

The preferred treatment criterion represents a relaxation of treatment parity. That is, every decision making system that achieves treatment parity also satisfies the preferred treat-

[1] Although it is quite possible that certain *individuals* from the group may not prefer the set that maximizes the benefit for the *group as a whole*.

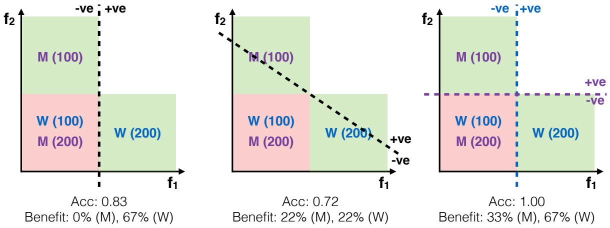

Figure 1: A fictitious decision making scenario involving two groups: men (M) and women (W). Feature f_1 (x-axis) is highly predictive for women whereas f_2 (y-axis) is highly predictive for men. Green (red) quadrants denote the positive (negative) class. Within each quadrant, the points are distributed uniformly and the numbers in parenthesis denote the number of subjects in that quadrant. The **left panel** shows the optimal classifier satisfying parity in treatment. This classifier leads to all the men getting classified as negative. The **middle panel** shows the optimal classifier satisfying parity in impact (in addition to parity in treatment). This classifier achieves impact parity by misclassifying women from positive class into negative class, and in the process, incurs a significant cost in terms of accuracy. The **right panel** shows a classifier consisting of group-conditional classifiers for men (purple) and women (blue). Both the classifiers satisfy the preferred treatment criterion since for each group, adopting the other group's classifier would lead to a smaller fraction of beneficial outcomes. Additionally, this group-conditional classifier is also a preferred impact classifier since both groups get more benefit as compared to the impact parity classifier and the overall accuracy is better.

ment condition, which implies (in theory) that the optimal decision accuracy that can be achieved under the preferred treatment condition is at least as high as the one achieved under treatment parity. Additionally, preferred treatment allows group-conditional decision making (not allowed by treatment parity), which is necessary to achieve high decision accuracy in scenarios when the predictive power of features varies greatly between different sensitive user groups, as shown in Figure 1.

While preferred treatment is a looser notion of fairness than treatment parity, it retains a core fairness property embodied in treatment parity, namely, *envy-freeness at the level of user groups*. Under preferred treatment, no group of users (*e.g.*, men or women, blacks or whites) would feel that they would be collectively better off by switching their group membership (*e.g.*, gender, race). Thus, preferred treatment decision making, despite allowing group-conditional decision making, is not vulnerable to being characterized as "reverse discrimination" against, or "affirmative action" for certain groups.

— **From Parity Impact to Preferred Impact:** To offer preferred impact, a decision making system needs to ensure that every sensitive attribute group (*e.g.*, men and women) *prefers* the set of decisions they receive over the set of decisions they would have received under the criterion of impact parity.

The preferred impact criterion represents a relaxation of impact parity. That is, every decision making system that achieves impact parity also satisfies the preferred impact condition, which implies (in theory) that the optimal decision accuracy that can be achieved under the preferred impact condition is at least as high as the one achieved under impact parity. Additionally, preferred impact allows disparity in benefits received by different groups, which may be justified in scenarios where insisting on impact parity would only lead to a reduction in the beneficial outcomes received by one or more groups, without necessarily improving them for any other group. In such scenarios, insisting on impact parity can additionally lead to a reduction in the decision accuracy, creating a case of tragedy of impact parity with a worse decision making all round, as shown in Figure 1.

While preferred impact is a looser notion of fairness compared to impact parity, by guaranteeing that every group receives *at least* as many beneficial outcomes as they would would have received under impact parity, it retains the core fairness gains in beneficial outcomes that discriminated groups would have achieved under the fairness criterion of impact parity.

Finally, we note that our preference-based fairness notions, while having may attractive properties, are not the most suitable notions of fairness in *all* scenarios. In certain cases, parity fairness may well be the eventual goal [1] and the more desirable notion.

2 Defining preference-based fairness

In this section, we first introduce two useful quality metrics—*utility* and *group benefit*—in the context of fairness in classification, then revisit parity-based fairness definitions in the light of these quality metrics, and finally formalize the two preference-based notions of fairness introduced in Section 1 from the perspective of the above metrics. For simplicity, we consider binary classification tasks, however, the definitions can be easily extended to m-ary classification.

Quality metrics in fair classification. In a fair (binary) classification task, one needs to find a mapping between the user feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and class labels $y \in \{-1, 1\}$, where $(\boldsymbol{x}, y)$ are drawn from an (unknown) distribution $f(\boldsymbol{x}, y)$. This is often achieved by finding a mapping function $\boldsymbol{\theta} : \mathbb{R}^d \to \mathbb{R}$ such that given a feature vector $\boldsymbol{x}$ with an unknown label y, the corresponding classifier predicts $\hat{y} = \text{sign}(\boldsymbol{\theta}(\boldsymbol{x}))$. However, this mapping function also needs to be *fair* with respect to the values of a user sensitive attribute $z \in \mathcal{Z} \subseteq \mathbb{Z}_{\geq 0}$ (*e.g.*, sex, race), which are drawn from an (unknown) distribution $f(z)$ and may be dependent of the feature vectors and class labels, *i.e.*, $f(\boldsymbol{x}, y, z) = f(\boldsymbol{x}, y|z) f(z) \neq f(\boldsymbol{x}, y) f(z)$.

Given the above problem setting, we introduce the following quality metrics, which we will use to define and compare different fairness notions:

Utility ($\mathcal{U}$) is defined as the overall profit obtained by the decision maker using the classifier. For instance, in a loan approval scenario, the decision maker is the bank that gives the loans and the utility can be the overall accuracy of the classifier, *i.e.*:

$$\mathcal{U}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x},y}[\mathbb{I}\{\text{sign}(\boldsymbol{\theta}(\boldsymbol{x})) = y\}],$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and the expectation is taken over $f(\boldsymbol{x}, y)$. It is in the decision maker's interest to use classifiers that maximize utility.

Group benefit ($\mathcal{B}_z$) are defined as the fraction of beneficial outcomes received by users sharing a certain value of the sensitive attribute z (*e.g.*, blacks, hispanics, whites). For example, in a loan approval scenario, the beneficial outcome for a user may be receiving the loan and the group benefit for each value of z can be defined as:

$$\mathcal{B}_z(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}|z}[\mathbb{I}\{\text{sign}(\boldsymbol{\theta}(\boldsymbol{x})) = 1\}], \tag{1}$$

where the expectation is taken over the conditional distribution $f(\boldsymbol{x}|z)$ and the bank offers a loan to a user if $\text{sign}(\boldsymbol{\theta}(\boldsymbol{x})) = 1$. In certain scenarios, as suggested by previous work [12, 15, 22], the group benefits can also be defined as the fraction of beneficial outcomes conditional on the true label of the user:

$$\mathcal{B}_z(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}|z,y=1}[\mathbb{I}\{\text{sign}(\boldsymbol{\theta}(\boldsymbol{x})) = 1\}].$$

Parity-based fairness. A number of recent studies [4, 9, 12, 14, 21, 22, 23] have considered a classifier to be fair if it satisfies the impact parity criterion:

$$\mathcal{B}_z(\boldsymbol{\theta}) = \mathcal{B}_{z'}(\boldsymbol{\theta}) \quad \text{for all } z, z' \in \mathcal{Z}. \tag{2}$$

Although not always explicitly sought, most of the above studies propose classifiers that also satisfy treatment parity in addition to impact parity, *i.e.*, they do not use the sensitive attribute z in the decision making process. However, some of them [4, 12, 14] do not satisfy treatment parity since they resort to group-conditional classifiers, *i.e.*, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z\}_{z \in \mathcal{Z}}$. In such case, we can rewrite the above parity condition as:

$$\mathcal{B}_z(\boldsymbol{\theta}_z) = \mathcal{B}_{z'}(\boldsymbol{\theta}_{z'}) \quad \text{for all } z, z' \in \mathcal{Z}. \tag{3}$$

Fairness beyond parity. Given the above quality metrics, we can now formalize the two preference-based fairness notions introduced in Section 1:

A classifier $\boldsymbol{\theta}$ resorting to group-conditional classifiers, *i.e.*, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z\}_{z \in \mathcal{Z}}$, is a **preferred treatment** classifier if each group sharing a sensitive attribute value z benefits more from its corresponding group-conditional classifier $\boldsymbol{\theta}_z$ than it would benefit if it would be classified by any of the other group-conditional classifiers $\boldsymbol{\theta}_{z'}$, *i.e.*,

$$\mathcal{B}_z(\boldsymbol{\theta}_z) \geq \mathcal{B}_z(\boldsymbol{\theta}_{z'}) \quad \text{for all } z, z' \in \mathcal{Z}. \tag{4}$$

Note that, if a classifier $\boldsymbol{\theta}$ does not use group-conditional classifiers, *i.e.*, $\boldsymbol{\theta}_z = \boldsymbol{\theta}$ for all $z \in \mathcal{Z}$, it will be always be a preferred treatment classifier. If, in addition, such classifier ensures impact parity, it is easy to show that its utility cannot be larger than a preferred treatment classifier consisting of group-conditional classifiers.

A classifier $\boldsymbol{\theta}$ offers **preferred impact** over a classifier $\boldsymbol{\theta}'$ ensuring impact parity if it achieves higher group benefit for each sensitive attribute value group, *i.e.*,

$$\mathcal{B}_z(\boldsymbol{\theta}) \geq \mathcal{B}_z(\boldsymbol{\theta}') \quad \text{for all } z \in \mathcal{Z}. \tag{5}$$

One can also rewrite the above condition for group-conditional classifiers, *i.e.*, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z\}_{z \in \mathcal{Z}}$ and $\boldsymbol{\theta}' = \{\boldsymbol{\theta}'_z\}_{z \in \mathcal{Z}}$, as follows:

$$\mathcal{B}_z(\boldsymbol{\theta}_z) \geq \mathcal{B}_z(\boldsymbol{\theta}'_z) \quad \text{for all } z \in \mathcal{Z}. \tag{6}$$

Note that, given any classifier $\boldsymbol{\theta}'$ ensuring impact parity, it is easy to show that there will always exist a preferred classifier $\boldsymbol{\theta}$ with equal or higher utility.

On individual-level preferences. Notice that preferred treatment and preferred impact notions are defined based on the group preferences, *i.e.*, whether a *group as a whole* prefers (or, gets more beneficial outcomes from) a given set of outcomes over another set. It is quite possible that a set of outcomes preferred by the group collectively is not preferred by certain *individuals* in the group. Consequently, one can extend these notions to account for individual preferences as well, *i.e.*, a set of outcomes is preferred over another if all the individuals in the group prefer it. In this paper, we focus on preferred treatment and preferred impact in the context of group preferences, and leave the case of individual preferences and its implications on the cost of achieving fairness to be explored thoroughly in a future study.

3 Training preferred classifiers

In this section, our goal is training preferred treatment and preferred impact group-conditional classifiers, *i.e.*, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_z\}_{z \in \mathcal{Z}}$, that maximize utility given a training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i, z_i)\}_{i=1}^{N}$, where $(\boldsymbol{x}_i, y_i, z_i) \sim f(\boldsymbol{x}, y, z)$. In both cases, we will assume that:[2] each group-conditional classifier is a linear boundary-based classifier, *i.e.*, $\boldsymbol{\theta}_z(\boldsymbol{x}) = \boldsymbol{\theta}_z^T \boldsymbol{x}$,

[2] Exploring the relaxations of these assumptions is a very interesting avenue for future work.

the utility function $\mathcal{U}$ is defined as the overall accuracy and the group benefit $\mathcal{B}_z$ for users sharing the sensitive attribute value z is defined as their average probability of being classified into the positive class (Eq. 1).

Preferred impact classifiers. Given a impact parity classifier with decision boundary parameters $\{\boldsymbol{\theta}'_z\}_{z \in \mathcal{Z}}$ and a convex loss function $\ell_{\boldsymbol{\theta}}$ of a linear classifier along with the regularization function $\Omega(.)$, one could think of finding the decision boundary parameters $\{\boldsymbol{\theta}_z\}_{z \in \mathcal{Z}}$ of a preferred impact classifier (Eq. 6) that maximizes utility by solving:

$$\begin{array}{ll} \underset{\{\boldsymbol{\theta}_z\}}{\min} & \frac{1}{N} \sum_{(\boldsymbol{x},y,z) \in \mathcal{D}} \ell_{\boldsymbol{\theta}_z}(\boldsymbol{x}, y) + \sum_{z \in \mathcal{Z}} \lambda_z \Omega(\boldsymbol{\theta}_z) \\ \text{s.t.} & \hat{\mathcal{B}}_z(\boldsymbol{\theta}_z) \geq \hat{\mathcal{B}}_z(\boldsymbol{\theta}'_z) \;\; \forall z \in \mathcal{Z}, \end{array} \tag{7}$$

where λ_z is the regularization strength for $\boldsymbol{\theta}_z$ and $\hat{\mathcal{B}}_z$ denotes the empirical group benefit. Note that the right hand side of the inequalities does not contain any variables, *i.e.*, the impact parity classifiers $\{\boldsymbol{\theta}'_z\}_{z \in \mathcal{Z}}$ are given.

Unfortunately, it is very challenging to solve the above problem since, for nontrivial linear-classifiers (*e.g.*, logistic regression, SVMs), the constraints are non-convex. To overcome this difficulty, we approximate the empirical benefit using a ramp (convex) function $r(x) = \max(0, x)$, *i.e.*,

$$\begin{array}{ll} \underset{\{\boldsymbol{\theta}_z\}}{\min} & \frac{1}{N} \sum_{(\boldsymbol{x},y,z) \in \mathcal{D}} \ell_{\boldsymbol{\theta}_z}(\boldsymbol{x}, y) + \sum_{z \in \mathcal{Z}} \lambda_z \Omega(\boldsymbol{\theta}_z) \\ \text{s.t.} & \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, \boldsymbol{\theta}_z^T \boldsymbol{x}) \geq \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, \boldsymbol{\theta}'^T_z \boldsymbol{x}) \quad \forall z \in \mathcal{Z}, \end{array} \tag{8}$$

where $\mathcal{D}_z = \{(\boldsymbol{x}_i, y_i, z_i) \in \mathcal{D} \,|\, z_i = z\}$ denotes the set of users with sensitive attribute value z. Eq. 8 is a disciplined convex-concave program (DCCP) for any convex regularizer $\Omega(\cdot)$ and can be efficiently solved using well-known heuristics [19]. The above formulation, for example, can be particularized for logistic regression classifier with L_2-norm regularizer by having $\ell_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = log(1 + exp(y\boldsymbol{\theta}^T \boldsymbol{x}))$ and $\Omega(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||^2$. One can similarly particularize the formulation for other convex boundary-based classifiers like squared loss, linear / non-linear SVMs, etc.

Preferred treatment classifiers. Using the definition of preferred treatment in Eq. 4, one can follow similar steps as preferred impact and find the optimal decision boundary parameters $\{\boldsymbol{\theta}_z\}_{z \in \mathcal{Z}}$ of a preferred treatment classifier as:

$$\begin{array}{ll} \underset{\{\boldsymbol{\theta}_z\}}{\min} & -\frac{1}{N} \sum_{(\boldsymbol{x},y,z) \in \mathcal{D}} \ell_{\boldsymbol{\theta}_z}(\boldsymbol{x}, y) + \sum_{z \in \mathcal{Z}} \lambda_z \Omega(\boldsymbol{\theta}_z) \\ \text{s.t.} & \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, \boldsymbol{\theta}_z^T \boldsymbol{x}) \geq \sum_{\boldsymbol{x} \in \mathcal{D}_z} \max(0, \boldsymbol{\theta}_{z'}^T \boldsymbol{x}) \; \forall z, z' \in \mathcal{Z}, \end{array} \tag{9}$$

which is also a disciplined convex-concave program (DCCP) for any convex regularizer $\Omega(\cdot)$ and can be efficiently solved. Here, note that unlike Eq. 8, both the left and right hand side of the inequalities contain optimization variables.

4 Evaluation and discussion

Here, we compare the performance of preferred treatment and impact classifiers against unconstrained, treatment parity and impact parity classifiers on the ProPublica COMPAS dataset [16]. The classification task is to predict whether a criminal defendant would recidivate within two years (negative class) or not (positive class). We use the same set of features as used by Zafar et al. [21] for training the classifiers. We designate race as the sensitive attribute and divide the subjects into two groups: blacks (group-0) and whites (group-1). The group benefits are computed as the fraction of subjects being classified into the positive class.

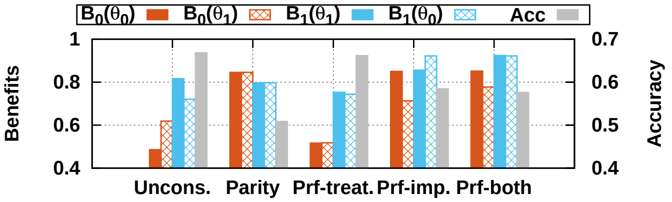

Figure 2: The figure shows the overall accuracy and the benefits received by the two groups for various classifiers. 'Prf-treat.', 'Prf-imp.', and 'Prf-both' respectively correspond to the classifiers satisfying preferred treatment, preferred impact, and both preferred treatment and impact criteria. Sensitive attribute values 0 and 1 denote blacks and whites, respectively. $\mathcal{B}_i(\boldsymbol{\theta}_j)$ denotes the benefits obtained by group i when using the classifier of group j. For the *Parity* case, we train just one classifier for both the groups, so the benefits do not change by adopting other group's classifier.

Next, we consider the following classifiers, which we train to maximize utility subject to the corresponding constraints:[3] An unconstrained (***Uncons***) classifier that resorts to group-conditional classifiers. It violates treatment parity and potentially violates impact parity as well. A ***parity*** classifier that does not use the sensitive attribute group information in the decision making, and is constrained to satisfy both treatment parity—its decisions do not change based on the users' sensitive attribute value as it does not resort to group-conditional classifiers—and impact parity—it ensures that the benefits for all groups are the same. We train this parity classifier using the methodology proposed by Zafar et al. [22]. And finally, we train a ***preferred treatment*** classifier (Eq. 9), a ***preferred impact*** classifier (Eq. 8), and a classifier (***preferred both***) which is both a preferred treatment as well as a preferred impact classifier. All the preferred classifiers consist of group-conditional classifiers.

The results, presented in Figure 2, show that the ***Uncons*** classifier, in addition to violating treatment parity (a separate classifier for each group) and impact parity (high disparity in group benefits), also violates the preferred treatment criterion (group-0 would benefit more by adopting group-1's classifier). On the other hand, the ***Parity*** classifier satisfies the treatment parity and impact parity but it does so at a large

[3] We use logistic regression classifiers with L_2-norm regularization.

cost in terms of accuracy, which is very close to that of a random classifier in this case.

The ***Preferred treatment*** classifier provides a much higher accuracy than the *Parity* classifier (on par with that of the *Uncons* classifier) while satisfying the preferred treatment criterion. However, it does not meet the preferred impact criterion. The ***Preferred impact*** classifier meets the preferred impact criterion with a larger drop in accuracy (which is still better than the parity classifier) but does not satisfy preferred treatment. Finally, the classifier satisfying both preferred treatment and preferred impact (***Preferred both***) leads to a further slight drop in terms of accuracy.

In summary, the results show that the preference-based notions of fairness can lead to significant performance gains as compared to parity-based notions of fairness. However, the precise performance gains may change depending on the underlying distribution of the dataset. Furthermore, we also note that splitting the datasets into sensitive attribute value groups for training group conditional classifiers may lead to degraded effectiveness of the empirical risk minimization procedure (since each classifier has fewer data points as compared to the parity classifier that is trained on the whole data). This problem may however be solved by gathering more data for each group. Finally, the ramp function-based proxies that we propose for training preferred classifiers are disciplined convex-concave programs (DCCP). While such programs can be efficiently solved using heuristic-based methods [19], unlike convex programs, the optimality of the DCCP solutions is not guaranteed. Addressing these issues would be interesting directions for the future work.

Acknowledgments. AW acknowledges support by the Alan Turing Institute under EPSRC grant EP/N510129/1, and by the Leverhulme Trust via the CFI.

References

[1] A. Altman. Discrimination. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016. https://plato.stanford.edu/archives/win2016/entries/discrimination/.

[2] S. Barocas and A. D. Selbst. Big Data's Disparate Impact. *California Law Review*, 2016.

[3] M. Berliant and W. Thomson. On the Fair Division of a Heterogeneous Commodity. *Journal of Mathematics Economics* , 1992.

[4] T. Calders and S. Verwer. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery*, 2010.

[5] A. Chouldechova. Fair Prediction with Disparate Impact:A Study of Bias in Recidivism Prediction Instruments. *arXiv preprint, arXiv:1610.07524*, 2016.

[6] Civil Rights Act. Civil Rights Act of 1964, Title VII, Equal Employment Opportunities, 1964.

[7] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic Decision Making and the Cost of Fairness. In *KDD*, 2017.

[8] C. Dwork, M. Hardt, T. Pitassi, and O. Reingold. Fairness Through Awareness. In *ITCSC*, 2012.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and Removing Disparate Impact. In *KDD*, 2015.

[10] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *arXiv:1609.07236*, 2016.

[11] G. Goh, A. Cotter, M. Gupta, and M. Friedlander. Satisfying Real-world Goals with Dataset Constraints. In *NIPS*, 2016.

[12] M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. In *NIPS*, 2016.

[13] F. Kamiran and T. Calders. Classification with No Discrimination by Preferential Sampling. In *Proceedings of Machine Learning conference of Belgium and The Netherlands*, 2010.

[14] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware Classifier with Prejudice Remover Regularizer. In *PADM*, 2011.

[15] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*, 2017.

[16] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. https://github.com/propublica/compas-analysis, 2016.

[17] B. T. Luong, S. Ruggieri, and F. Turini. kNN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *KDD*, 2011.

[18] C. Muñoz, M. Smith, and D. Patil. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. *Executive Office of the President. The White House.*, 2016.

[19] X. Shen, S. Diamond, Y. Gu, and S. Boyd. Disciplined Convex-Concave Programming. *arXiv:1604.02639*, 2016.

[20] H. R. Varian. Equity, Envy, and Efficiency. *Journal of Economic Theory*, 1974.

[21] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*, 2017.

[22] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*, 2017.

[23] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning Fair Representations. In *ICML*, 2013.