Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning

Nina Grgić-Hlača[1], Muhammad Bilal Zafar[1], Krishna P. Gummadi[1], Adrian Weller[2,3,4]*

[1]Max Planck Institute for Software Systems (MPI-SWS), Germany
[2]University of Cambridge, UK
[3]Alan Turing Institute, UK
[4]Leverhulme Centre for the Future of Intelligence, UK
{nghlaca, mzafar, gummadi}@mpi-sws.org, adrian.weller@eng.cam.ac.uk

Abstract

With widespread use of machine learning methods in numerous domains involving humans, several studies have raised questions about the potential for unfairness towards certain individuals or groups. A number of recent works have proposed methods to measure and eliminate unfairness from machine learning models. However, most of this work has focused on only one dimension of fair decision making: *distributive fairness*, *i.e.*, the fairness of the decision *outcomes*. In this work, we leverage the rich literature on organizational justice and focus on another dimension of fair decision making: *procedural fairness*, *i.e.*, the fairness of the decision making *process*. We propose measures for procedural fairness that consider the input features used in the decision process, and evaluate the moral judgments of humans regarding the use of these features. We operationalize these measures on two real world datasets using human surveys on the Amazon Mechanical Turk (AMT) platform, demonstrating that our measures capture important properties of procedurally fair decision making. We provide fast submodular mechanisms to optimize the tradeoff between procedural fairness and prediction accuracy. On our datasets, we observe empirically that procedural fairness may be achieved with little cost to outcome fairness, but that some loss of accuracy is unavoidable.

1 Introduction

As machine learning methods are increasingly being used in decision making scenarios that affect human lives (such as credit risk assessments and recidivism risk prediction), there are growing concerns about the *fairness* of such decision making. These concerns have spawned much recent research on detecting and avoiding unfairness in decision making (Dwork et al. 2012; Feldman et al. 2015; Kamiran and Calders 2010; Luong, Ruggieri, and Turini 2011; Pedreschi, Ruggieri, and Turini 2008; Zafar et al. 2017b; Zemel et al. 2013).

In this work, we revisit the foundational notions of fairness that underlie these studies. We begin by observing that fairness concerns about decision making are *multi-dimensional*. Specifically, we leverage the rich literature on *organizational justice* (Greenberg 1987) to distinguish between two primary categories (types) of fairness concerns, namely **distributive** and **procedural** fairness. Distributive fairness refers to the fairness of the *outcomes (ends)* of decision making, while procedural fairness refers to the fairness of the decision making *processes (means)* that lead to the outcomes. In the rest of this paper, we use "distributive fairness" and "outcome fairness" interchangeably. We similarly use the terms "procedural fairness" and "process fairness" interchangeably.

To date, most works on fair learning have focused on achieving a fair distribution of decision outcomes, with little attention to the decision processes generating the outcomes. These works are inspired by various anti-discrimination laws that focus on the relationships between certain sensitive attributes (*e.g.*, gender or race) and decision outcomes (Civil Rights Act 1964; Barocas and Selbst 2016). For instance, the notions of "individual fairness" (Dwork et al. 2012), "situational testing" (Luong, Ruggieri, and Turini 2011), and "disparate treatment" (Zafar et al. 2017b) consider individuals who belong to different sensitive attribute groups (*e.g.*, males and females), yet share similar non-sensitive features (qualifications), and require them to receive *similar decision outcomes*. Similarly, the notions of "group fairness" (Zemel et al. 2013) and "disparate impact" (Zafar et al. 2017b) are based on requiring that different sensitive attribute groups receive *beneficial decision outcomes in similar proportions*. Finally, the notions of "disparate mistreatment" (Zafar et al. 2017a) and "equal opportunity" (Hardt, Price, and Srebro 2016) aim to achieve *similar rates of errors in decision outcomes* for different sensitive attribute groups.

In this paper, we propose notions of procedural (rather than distributive) fairness, based on which input features are used in the decision process and how including or excluding the features would affect outcomes. While existing fair learning mechanisms can efficiently leverage input features and their correlations with the sensitive attributes in order to resolve indirect discrimination and achieve distributively fair outcomes, they overlook several important considerations which are addressed by procedural fairness. For example, these considerations include whether or not the perceived fairness of using an individual's feature in the deci-

*Adrian Weller acknowledges the support of the David MacKay Newton research fellowship at Darwin College, the Alan Turing Institute under EPSRC grant EP/N510129/1, and the Leverhulme Trust via the CFI.

sion making process is affected by the following:

1. Feature volitionality: Does the feature represent the result of volitional (*i.e.*, voluntarily chosen) decisions made by the individual (*e.g.*, number of prior offenses); or rather is it the result of circumstances beyond their control (*e.g.*, age or race) (Beahrs 1991)?

2. Feature reliability: How reliably can a feature be assessed (*e.g.*, in credit assessments, opinions towards bankruptcy may be harder to reliably assess than number of prior bankruptcies) (Trankell 1972)?

3. Feature privacy: Does use of the feature give rise to a violation of the individual's privacy (GDPR 2016)?

4. Feature relevance: Is the feature causally related or not to the decision outcomes (Kilbertus et al. 2017; Kusner et al. 2017)?

The above procedural fairness considerations reflect an understanding of the potential harmful impact on society by unacceptable use of input features in a decision process. Note that these considerations depend on background knowledge about the input features and societal context that is often not captured in the data at hand. The challenge we then face is: *how should one account for procedural fairness in decision making?*

In this work, we explore a novel approach to procedurally fair learning. Our key insight, inspired by the work of Yaari and Bar-Hillel (1984) is to rely on humans' *moral judgments or instincts* about the fairness of using input features in a decision making context. As we shall show, humans (even lay users) exhibit a *moral sense* for whether or not it is fair to use a feature in a decision making scenario, that captures several of the procedural fairness considerations listed above. Their collective opinions – rooted in prevailing cultural or social norms, political beliefs, and legal regulations – reflect societal consensus on the desirability of using particular features.[1] These opinions (implicitly) provide the missing background knowledge needed to evaluate the fair use of input features in a given decision process.

In this paper, we propose and evaluate *measures and mechanisms* for procedurally fair learning. Our measures go beyond binary legal specifications – which mark each feature as either sensitive or not – by examining a scalar measure of the extent to which people judge each input feature to be (un)fair to use. Our measures are designed to avoid the traditional pitfalls of relying on human judgments. Specifically, we observe that while humans tend to have a good understanding of feature volitionality, reliability, privacy, and relevance in a decision making scenario, they tend to be bad at predicting what impact using a feature might have on the decision outcomes (Agan and Starr 2016). In fact, the impact of input features on outcomes is best assessed from the data itself. Accordingly, our measures explicitly seek to incorporate people's judgments of fairness of using a feature, conditioned on its impact on outcomes.

[1] Previous work has demonstrated the importance and validity of measuring perception of normative phenomena via human surveys, for example language impartiality (Zafar, Gummadi, and Danescu-Niculescu-Mizil 2016) and politeness (Danescu-Niculescu-Mizil et al. 2013).

We highlight the following contributions of our paper:

- We introduce three scalar measures of procedural fairness. In contrast to existing distributive fairness measures that are centered around decision outcomes, our measures account for fairness concerns about the use of input features in the process of decision making.
- We operationalize the measures by assembling user judgments of fairness of features in the context of recidivism risk estimation using the ProPublica COMPAS dataset (Larson et al. 2016). We also analyze the New York Police stop-question-and-frisk (SQF) dataset (SQF Dataset 2017), in the context of prediction of illegal weapon possession.
- We model the tradeoff between procedural fairness and the accuracy of a classifier as constrained submodular optimization problems over the set of features. We select subsets of features that optimize for accuracy and procedural fairness using fast and scalable methods for submodular optimization with submodular constraints, and demonstrate good performance empirically.
- On the datasets we considered, our results suggest that high procedural fairness, perhaps surprisingly, leads also to high distributive fairness, but that some loss of accuracy is unavoidable.

Related work in other disciplines. In moral philosophy (Blackburn 2003): a *deontological* approach considers certain moral truths to be absolute regardless of situation or outcome, which corresponds well with our notion of process fairness. In contrast, a *teleological* or *utilitarian* approach focuses on the outcomes, which corresponds well with the notion of outcome fairness.

Prior work in economics, law, and political science distinguishes between *direct* and *indirect* discrimination, suggesting that the "wrong" of direct discrimination (which we identify with violating process fairness) should be distinguished from the "wrong" of indirect discrimination (which we identify with violating outcome fairness) (Altman 2016).

2 Defining process fairness

We consider a classifier $\mathcal{C}$ trained using a subset of features $\mathcal{F}$ from a set $\bar{\mathcal{F}}$ of all possible features. We define the *process fairness* of $\mathcal{C}$ to be the fraction of all users who consider the use of every feature in $\mathcal{F}$ to be fair. We may use the phrase "process fairness of a classifier" interchangeably with "process fairness of the set of features used by the classifier".

Our process fairness definition relies critically on users' judgments about the use of individual features when making decisions. A user's judgment about a feature may change after learning how the use of the feature impacts decision outcomes. Some effects on decision outcomes are considered desirable, and might increase the perceived fairness of the feature. For instance, a user who initially considered a feature unfair for predicting recidivism risk might change their mind and deem the feature fair to use after learning that using the feature significantly improves the accuracy of prediction. On the other hand, other effects are considered undesirable, and may subsequently decrease the perceived fairness. For example, a user might change their mind after learning that using a feature would increase disparity in decision outcomes for different sensitive attribute groups (*e.g.*, men vs. women, white vs. non-white people).

In this paper, we focus on how the perceived fairness of a feature is affected by additional knowledge about a desirable effect: an increase in accuracy, and an undesirable effect: an increase in disparity in decision outcomes. Accordingly, we define three measures of process fairness: *feature-apriori fairness*, *feature-accuracy fairness* and *feature-disparity fairness*. Let $\mathcal{U}$ denote the set of all queried individuals ("users").

Feature-apriori fairness. For a given feature $f \in \bar{\mathcal{F}}$, let $\mathcal{U}_f \subseteq \mathcal{U}$ denote the set of all users who consider the feature f fair to use without *a priori* knowledge of how its usage affects outcomes. For a given set of features $\mathcal{F} \subseteq \bar{\mathcal{F}}$, we define the *feature-apriori fairness* of $\mathcal{F}$ as the fraction of users who consider *all* of the features $f \in \mathcal{F}$ fair. More formally,

$$\textit{feature-apriori fairness}(\mathcal{F}) := \frac{|\bigcap_{f\in\mathcal{F}} \mathcal{U}_f|}{|\mathcal{U}|}. \quad (1)$$

Feature-accuracy fairness. Let $\mathcal{U}_f^A \subseteq \mathcal{U}$ denote the set of all users who consider the feature f fair to use *if it increases the accuracy* of a designated classifier. Given a set of features $\mathcal{F} \subseteq \bar{\mathcal{F}}$, we define:

$$\textit{feature-accuracy fairness}(\mathcal{F}) := \frac{|\bigcap_{f\in\mathcal{F}} \mathcal{A}(\mathcal{U}_f, \mathcal{U}_f^A)|}{|\mathcal{U}|},$$

$$\mathcal{A}(\mathcal{U}_f, \mathcal{U}_f^A) = \begin{cases} \mathcal{U}_f^A, & \text{if } f \text{ increases accuracy} \\ \mathcal{U}_f, & \text{otherwise.} \end{cases} \quad (2)$$

Feature-disparity fairness. Let $\mathcal{U}_f^D \subseteq \mathcal{U}$ denote the set of all users who consider the feature f fair to use *even if it increases disparity in outcomes* of a designated classifier. In this paper, we use disparate mistreatment (Zafar et al. 2017a) as a measure of disparity in outcomes (see Section 6), but other measures such as disparate impact (Zafar et al. 2017b) could be used as well. Given a set of features $\mathcal{F} \subseteq \bar{\mathcal{F}}$, we define:

$$\textit{feature-disparity fairness}(\mathcal{F}) := \frac{|\bigcap_{f\in\mathcal{F}} \mathcal{D}(\mathcal{U}_f, \mathcal{U}_f^D)|}{|\mathcal{U}|},$$

$$\mathcal{D}(\mathcal{U}_f, \mathcal{U}_f^D) = \begin{cases} \mathcal{U}_f^D, & \text{if } f \text{ increases disparity} \\ \mathcal{U}_f, & \text{otherwise.} \end{cases} \quad (3)$$

Following similar reasoning as for feature-accuracy and feature-disparity fairness, one could define measures of process fairness that capture other desirable and undesirable effects of using features.

Unfairness measures. Our fairness measures are set functions: they take a subset of features $\mathcal{F} \subseteq \bar{\mathcal{F}}$ and return a real number as a fairness value between 0 (completely unfair) and 1 (completely fair). For each of our three measures of process fairness, we define corresponding measures of *unfairness*, each defined as 1 minus the respective fairness measure.

Details of implementation. For definitions (2) and (3), the classifier $\mathcal{C}$ must be specified in order to determine the $\mathcal{A}$ and $\mathcal{D}$ condition functions. Furthermore, to determine if using a feature f increases accuracy or disparity, we must specify: (i) a minimum threshold level ϵ for the increase; and (ii) the base set of features with respect to which f increases the property. For (i), we set $\epsilon = 5\%$ of the full range of values realized by the null classifier and the classifier which uses all features $\bar{\mathcal{F}}$. For (ii) we used the empty set of features as the base set. Other reasonable choices yield qualitatively similar results on our datasets.

Properties of (un)fairness measures. In Proposition 1 below, we show key properties of our measures which will be critical to enable us in Section 4 to develop scalable optimization methods for establishing fairness-accuracy tradeoffs. We first need the following definitions. Let $\bar{\mathcal{F}}$ be a finite set. Let g be a set function $g : 2^{\bar{\mathcal{F}}} \to \mathbb{R}$, where $2^{\bar{\mathcal{F}}}$ denotes the power set of $\bar{\mathcal{F}}$.

Definition 1 *The function g is supermodular if for all $\mathcal{F}_A \subseteq \mathcal{F}_B \subset \bar{\mathcal{F}}, f \in \bar{\mathcal{F}} \setminus \mathcal{F}_B$,*

$$g(\mathcal{F}_A \cup \{f\}) - g(\mathcal{F}_A) \leq g(\mathcal{F}_B \cup \{f\}) - f(\mathcal{F}_B). \quad (4)$$

Intuitively, a set function is supermodular if it exhibits increasing marginal gains. A function is *submodular* if and only if its negative is supermodular.

Definition 2 *The function g is non-increasing monotone if:*

$$g(\mathcal{F} \cup \{f\}) - g(\mathcal{F}) \leq 0, \ \forall \mathcal{F} \subseteq \bar{\mathcal{F}}, f \in \bar{\mathcal{F}} \setminus \mathcal{F}. \quad (5)$$

Proposition 1 *All three measures of process fairness (feature-apriori, feature-accuracy and feature-disparity) are monotone non-increasing supermodular set functions with respect to features. Equivalently, the respective unfairness measures are monotone non-decreasing submodular functions.*

Proof Let g be any of the three measures of process fairness. We must show that, for any two $\mathcal{F}_A, \mathcal{F}_B$ such that $\mathcal{F}_A \subseteq \mathcal{F}_B \subset \bar{\mathcal{F}}$, and for any $f \in \bar{\mathcal{F}} \setminus \mathcal{F}_B$, inequality (4) is satisfied.

For ease of exposition, assume feature-apriori fairness, and let $\mathcal{F}^\cap$ denote $\bigcap_{f\in\mathcal{F}} \mathcal{U}_f$. Since $|\mathcal{U}| > 0$ is a constant, to show that inequality (4) holds, it is sufficient to show that:

$$|\mathcal{F}_A^\cap| - |\mathcal{F}_A^\cap \cap \{f\}^\cap| \geq |\mathcal{F}_B^\cap| - |\mathcal{F}_B^\cap \cap \{f\}^\cap|, \quad (6)$$

or, since for any X and v, $X = (X \cap \{v\}) \cup (X \setminus \{v\})$, that:

$$|\mathcal{F}_A^\cap \setminus \{f\}^\cap| \geq |\mathcal{F}_B^\cap \setminus \{f\}^\cap|. \quad (7)$$

Since $\mathcal{F}_A \subseteq \mathcal{F}_B$, it must be the case that $\mathcal{F}_B^\cap \subseteq \mathcal{F}_A^\cap$. It follows that $\mathcal{F}_B^\cap \setminus \{f\}^\cap \subseteq \mathcal{F}_A^\cap \setminus \{f\}^\cap$. Hence, inequality (7) holds and g is supermodular.

For any set $\mathcal{F} \subseteq \bar{\mathcal{F}}, f \in \bar{\mathcal{F}} \setminus \mathcal{F}$, it holds that $|\mathcal{F}^\cap \cap \{f\}^\cap| \leq |\mathcal{F}^\cap|$, hence inequality (5) also holds and the result follows. ■

3 Measuring process fairness

Here we apply the measures of process fairness defined in Section 2 to the ProPublica COMPAS dataset and the NYPD SQF dataset.

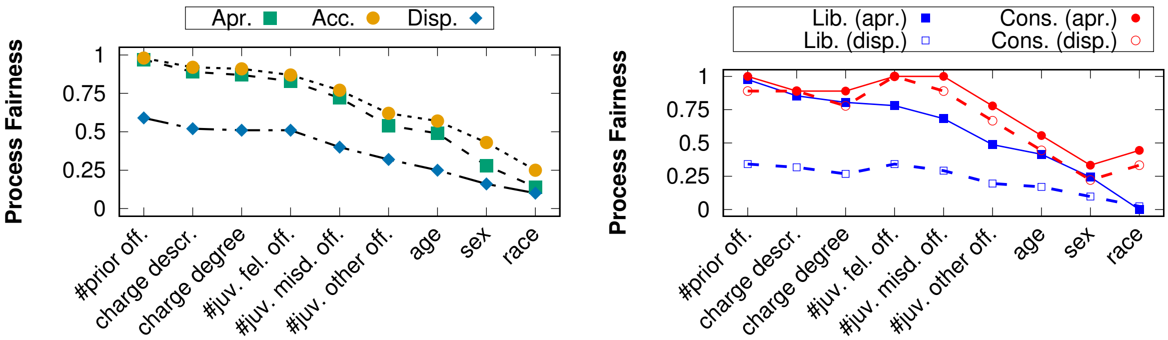

Figure 1: ProPublica COMPAS dataset. [**Left**] – Feature-apriori (apr.), feature-accuracy (acc.) and feature-disparity (disp.) fairness measured using judgments of AMT workers. [**Right**] – Fairness measured using judgments of very liberal (lib.) and very conservative (cons.) AMT workers.

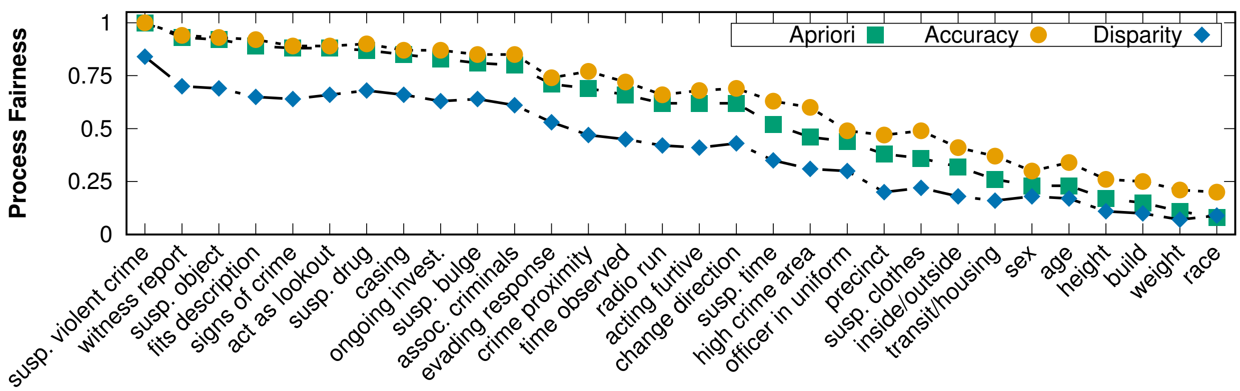

Figure 2: NYPD SQF dataset. Feature-apriori, feature-accuracy and feature-disparity fairness measured using judgments of AMT workers.

Our goal here is to utilize human judgments of fairness, in order to capture background knowledge (missing in the dataset) about a feature's volitionality, relevance, and other properties. In this paper, we survey users of Amazon Mechanical Turk (AMT) platform to gather human judgments of fairness. However, we note that such judgments can be gathered from any other group of people, ranging from crowd workers to domain experts. We do *not* claim that AMT users are the right group to obtain these judgments from. We first describe how we surveyed users to gather these judgments.

3.1 Gathering human judgments

For gathering human judgments, we use the Amazon Mechanical Turk (AMT) platform where users (or workers) can volunteer to perform a wide range of online tasks for pay (Buhrmester, Kwang, and Gosling 2011; Mason and Suri 2012).

For each of the datasets, we describe the prediction task, and then present one feature at a time, asking the user each time to respond with yes or no to three questions: **Q. 1:** Do you believe it is fair or unfair to use this information? **Q. 2:** Do you believe it is fair or unfair to use this information, if it *increases the accuracy* of the prediction? **Q. 3:** Do you believe it is fair or unfair to use this information, if it makes one group of people (*e.g.*, African American people) *more likely to be falsely predicted as having a higher risk of recidivism* than another group of people (*e.g.*, white people)?

We intentionally did not define *fair* in our questions in order to gather users' intuitive sense of fairness. As we discuss in Section 1, users' fairness judgments reflect many complex considerations.

For a given dataset, we gather responses to the above questions from 200 different AMT workers (that is, each feature is judged by 200 different workers). Since the tasks we consider relate to the U.S. criminal justice system, we only recruited workers who are from the U.S. To ensure the quality of the judgments, we only recruited AMT *master* workers who have a reputation on the AMT platform for performing their tasks reliably (The Mechanical Turk Blog 2011). Additionally, we filtered out the fairness judgments from a small fraction (less than 5%) of users, who provided outlier (anomalous) responses: (i) marking a feature as unfair to use apriori to knowing its impact, but marking it as fair when it increases disparity and (ii) marking a feature as fair to use apriori to knowing its impact, but marking it as unfair when it increases accuracy. Below, we describe the ProPublica COMPAS and NYPD SQF datasets.

ProPublica COMPAS dataset. The ProPublica COMPAS dataset (Larson et al. 2016) relates to *recidivism risk prediction* (predicting if a criminal defendant will commit an offense within a certain future time). The dataset is gathered by ProPublica (Larson et al. 2016), with information on all criminal defendants who were subject to screening by COMPAS, a commercial recidivism risk assessment tool, in Broward County, Florida, in 2013-2014. Our set of features was the following: "number of prior criminal offenses", "arrest charge description" (*e.g.*, grand theft, possession of drugs), "charge degree" (misdemeanor or felony), "number of juvenile felony offenses", "juvenile misdemeanor offenses", "other juvenile offenses", "age" of the defendant, "sex" of the defendant and "race" of the defendant. A subset of these features was considered in recent studies related to racial biases in recidivism risk prediction (Flores, Lowenkamp, and Bechtel 2016; Zafar et al. 2017a). The dataset also contains information on whether the defendant actually recidivated or not.

NYPD SQF dataset. We also gathered responses from AMT workers for a dataset related to New York Police Department's Stop-Question-and-Frisk (NYPD SQF) program (SQF Dataset 2017), where police officers stop and investigate civilians on the suspicion of being involved in a criminal activity. The dataset is publicly available (SQF Dataset 2017) and has been studied by various prior works in the context of outcome fairness (Goel, Rao, and Shroff 2015). It contains data on all of the stops made by police officers, including the accompanying circumstances and reasons for the stop. In a recent study, Goel, Rao, and Shroff (2015) designed a learning task based on appropriate features present in the datasets to predict whether a person is carrying an illegal weapon or not. We used the same prediction task and a similar set of 30 features as considered by Goel, Rao, and Shroff (2015), on the NYPD SQF data for 2011. The class distribution of the SQF dataset is highly skewed, with 97% of the instances belonging to the positive class. A classifier trained on such a dataset predicts all points as positive, while achieving an accuracy of 0.97. Therefore, we subsample the dataset to have equal class distribution.

3.2 Analyzing human judgments of fairness

For each feature, we computed the fraction of AMT workers who considered it fair under each of the questions Q. 1, 2 and 3, which correspond to notions of feature-apriori, feature-accuracy and feature-disparity fairness respectively. The results are shown in Figures 1 [Left] and 2, for the ProPublica COMPAS and NYPD SQF datasets respectively. Responses varied significantly across features, while the ranking of features was consistent across the three measures. As expected, compared to feature-apriori fairness, feature-accuracy fairness is higher and feature-disparity fairness is significantly lower.

ProPublica COMPAS dataset. In Figure 1 [Left], we see that the features from the ProPublica COMPAS dataset neatly fall into three subsets, with declining levels of reported fairness. The first subset consists of features which are *directly related* to the issue at hand, such as the nature of the current charge. Next are *distantly related* features which provide information about the defendant's past record as a juvenile. The third set contains features which appear *unrelated*, such as sex and race. With this perspective, the users' responses may appear reasonable. In addition, note that the first two (most fair) sets contain *volitional* features, that is they relate to actions which the defendant chose to take, and hence might reasonably be considered predictive of the defendant's future actions; whereas the third (most unfair) set comprises features which are *physiological* and beyond the defendant's control. The third set is often considered protected by law (Civil Rights Act 1964). However, our results provide a more nuanced, scalar view of the judged fairness of all features.

NYPD SQF dataset. In the NYPD SQF dataset, we observe similar trends as in the ProPublica COMPAS dataset: features that are *directly related* to the issue at hand, such as "suspicion of engaging in a violent crime" ("susp. crime"), are rated as more fair than *distantly related* features, like "acting furtively", which are in turn rated as more fair than *unrelated* features, such as "sex" and "race". Similar conclusions as for the COMPAS dataset can be made about *volitional* and *physiological* features, as well, even though there are some exceptions, since "fitting a relevant description" is a *directly related*, but *physiological* feature.

Dependence on population demographics. For each AMT worker, we also gathered information on gender, race and political stance. Results were qualitatively similar across gender and race. However, "very liberal" and "very conservative" workers responded differently, as shown in Figure 1 [Right]. Conservative users rated the features as more fair than liberal users. Further, if a feature increased disparity, liberal users decreased their perceived fairness substantially more than conservative ones. This suggests that liberal users may be more sensitive to disparate outcomes, which is consistent with literature in the social sciences indicating that different political views may relate to different "moral foundations" (Graham et al. 2012).

4 Training procedurally fair classifiers

Thus far, we have used human judgments to quantify process fairness of each of the *individual* features in the ProPublica COMPAS and NYPD SQF datasets. Here, we begin to examine the process fairness of *sets of features* and their corresponding classifiers, using the definitions from Section 2. While excluding features deemed highly unfair from a classifier's inputs will increase its process fairness, it may lead to significantly lower prediction accuracy. We empirically analyze this tradeoff between process fairness and accuracy.

Classifier. Throughout this paper, we always use logistic regression with L2-regularization.[2] We chose this classifier for

[2]The classifier is implemented with the Python Scikit-learn package (Pedregosa et al. 2011). For all reported results, we randomly split the data into 50%/50% train/test folds 5 times and report average statistics. We use a regularization parameter of 1, as in (Kamishima et al. 2012). Other reasonable choices yielded similar results.

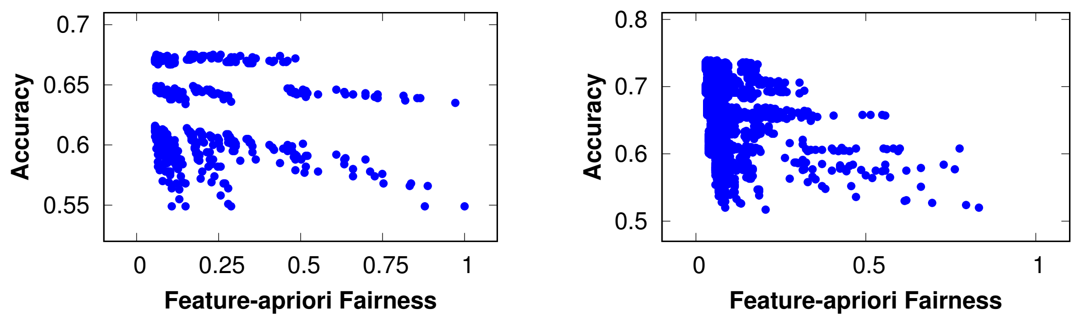

Figure 3: Tradeoffs between accuracy and feature-apriori fairness. The plots show the feature-apriori fairness (x-axis) and accuracy (y-axis) of classifiers trained on [**Left**] all $2^9 = 512$ subsets of the 9 features from the ProPublica COMPAS dataset, and [**Right**] a random sample of 5000 out of $2^{16} = 65,536$ different classifiers, trained on the 16 most informative features from the NYPD SQF dataset.

two reasons: (i) it has been used frequently in earlier works in this area (*e.g.*, (Goel, Rao, and Shroff 2015)); and (ii) it has attractive properties which will facilitate our approach to scalable optimization. We next explain these attractive properties.

For a given set of features $\mathcal{S} \subseteq \bar{\mathcal{F}}$ (where $\bar{\mathcal{F}}$ represents all the features that are available in the dataset under consideration), let $\text{acc}(\mathcal{S})$ be the set function given by the accuracy of our chosen classifier trained on the feature set $\mathcal{S}$. Intuitively, as features are added to $\mathcal{S}$, we expect that $\text{acc}(\mathcal{S})$ will rise but with diminishing returns. That is, we might expect that $\text{acc}(\mathcal{S})$ is *monotonically nondecreasing* and *submodular*. In practice, these properties do not always hold exactly (*e.g.*, due to curse of dimensionality). By leveraging the connection between strong convexity and submodularity, Elenberg et al. (2016) showed that $\text{acc}(\mathcal{S})$ is *weakly submodular*, which still enables fast approximate optimization.

Accuracy-fairness tradeoff. For the ProPublica COMPAS dataset, we train $2^9 = 512$ different classifiers – one for each possible subset of the 9 features present in the dataset. However, our NYPD SQF dataset has 30 features, which would lead to training 1.1 billion different classifiers, which is unworkable. Hence we selected a subset of the 16 most informative features using L1 feature selection (Tibshirani 1994),[3] and trained all $2^{16} = 65,536$ different classifiers.

Figure 3 shows plots of feature-apriori process fairness against accuracy (results for feature-accuracy and feature-disparity fairness are similar, omitted due to space constraints). Feature sets that correspond to high fairness (represented by points at the right side of the figures) come at the cost of accuracy. This effect is visible for both datasets, but it is even clearer for the NYPD SQF dataset in Figure 3 [Right].

We establish the optimal tradeoff in fair feature selection by finding solutions that lie along the upper envelope of points shown in Figure 3. Specifically, the key challenge lies in selecting a subset of features that either (1) optimize for accuracy, given a desired fairness threshold; or (2) optimize for fairness, given a desired accuracy threshold.[4] More formally:

(1) Maximizing accuracy under (un)fairness constraints: Consider a dataset $\mathcal{D}_{i=1}^{N}$ consisting of N records, each with a corresponding set of features $\bar{\mathcal{F}}$. For each record $\mathcal{D}_i$, let $y_i \in \{-1, 1\}$ be the decision variable. Assume that $\mathcal{D}^{\mathcal{S}}$, where $\mathcal{S} \subseteq \bar{\mathcal{F}}$, denotes the part of the dataset where for all records, only a subset $\mathcal{S}$ of all features $\bar{\mathcal{F}}$ is selected. Given this information, one can formulate the problem of training the most accurate classifier subject to process unfairness constraints as:

$$\begin{aligned} \underset{S \subseteq \bar{\mathcal{F}}}{\text{maximize}} \quad & acc(\mathcal{S}) \\ \text{subject to} \quad & unf(\mathcal{S}) \leq t, \end{aligned} \tag{8}$$

where $\text{acc}(\mathcal{S})$ and $\text{unf}(\mathcal{S})$ are set functions of $\mathcal{S} \subseteq \bar{\mathcal{F}}$, denoting the accuracy and unfairness of the corresponding classifiers. t is a desired threshold, specifying the maximum level of unfairness (minimum level of fairness) that is tolerable. For logistic regression (and linear classifiers in general), the accuracy, $\text{acc}(\mathcal{S})$, of a feature set $\mathcal{S}$, is computed as: $\frac{1}{N}\sum_{i=1}^{N} 1(sign(<\ \boldsymbol{\theta}^*, \mathcal{D}_i^{\mathcal{S}}\ >) = y_i)$,[5] where "$<>$" represents a dot product and $\boldsymbol{\theta}^*$ represents the optimal decision boundary parameters learned though empirical risk minimization. For logistic regression classifiers with L2-regularization, the optimal decision boundary $\boldsymbol{\theta}^*$ can be found by solving the following maximum likelihood

[3]We trained a logistic regression classifier with L1-regularization on all 30 features and selected the 16 features with highest absolute weights from the weight vector. These 16 features also cover the range of process fairness well, as seen in Figure 3 [Right].

[4]Similar problems have previously been studied in the context of feature selection (Iyer and Bilmes 2012), sensor placement (Krause, Singh, and Guestrin 2008), diversification (Ahmed, Dickerson, and Fuge 2017; Ashkan et al. 2015), document summarization (Lin and Bilmes 2011) and data subset selection (Lin and Bilmes 2009).

[5]To accommodate the classifier bias (or the intercept) term, we assume that all records in the dataset are padded with a dummy feature with constant value 1.

problem: $\boldsymbol{\theta}^* = \text{argmin}_{\boldsymbol{\theta}} - \sum_{i=1}^{N} \log p(y_i|\mathcal{D}_i^{\mathcal{S}}, \boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_2$, where $p(y_i = 1|\mathcal{D}_i^{\mathcal{S}}, \boldsymbol{\theta}) = \frac{1}{1+e^{-<\boldsymbol{\theta}, \mathcal{D}_i^{\mathcal{S}}>}}$, and $\lambda \in \mathbb{R}^+$ specifies the regularization strength (see footnote 2). In our experiments, we use $\lambda = 1$. The (un)fairness $\text{unf}(\mathcal{S})$ is defined according to three different notions (feature-apriori, feature-accuracy and feature disparity) presented in Section 2.

The optimization problem (8) can be solved rapidly provided that accuracy and unfairness are monotone and submodular set functions of $\mathcal{S}$. In this scenario, the above optimization formulation matches the canonical form of the submodular cost submodular knapsack (SCSK) problem (Iyer and Bilmes 2013), for which rapid approximate solutions have been proposed. Specifically, the iterated submodular-cost knapsack (ISK) algorithm proposed in Iyer and Bilmes (2013) offers $\left[1 - e^{-1}, \frac{K_{unf}}{1+(K_{unf}-1)(1-\kappa_{unf})}\right]$ performance bounds (on acc and unf respectively), where $K_{unf} = \max\{|\mathcal{S}| : unf(\mathcal{S}) \leq t\}$ and $\kappa_{unf} = 1 - \min_{f \in \mathcal{S}} \frac{unf(\mathcal{S}) - unf(\mathcal{S} \setminus \{f\})}{unf(\{f\})}$ is the total curvature of unf.

We have already showed in Section 2 that unfairness is a monotone submodular set function of $\mathcal{S}$. As discussed near the beginning of this section, Elenberg et al. (2016) showed that the accuracy of our classifier (L2-regularized logistic regression) is weakly submodular. Since accuracy is neither strictly monotone nor submodular, while solving (8) using the ISK algorithm (Iyer and Bilmes 2013), the performance bounds provided by Iyer and Bilmes (2013) need not hold precisely. However, as our empirical results on two different datasets in the next section show, in practice, solutions to (8) yielded by these algorithms are very close to the optimum.

(2) Minimizing unfairness under accuracy constraints: If instead one would like to achieve the least unfair (most fair) solution under a given accuracy performance constraint,[6] the tradeoff can alternatively be considered as follows:

$$\begin{aligned} \underset{S \subseteq \bar{\mathcal{F}}}{\text{minimize}} \quad & unf(\mathcal{S}) \\ \text{subject to} \quad & acc(\mathcal{S}) \geq t. \end{aligned} \tag{9}$$

Assuming $\text{unf}(\mathcal{S})$ and $\text{acc}(\mathcal{S})$ to be submodular set functions, problem (9) matches the canonical form of submodular cost submodular cover (SCSC) problem (Iyer and Bilmes 2013) and can be solved by using the iterated submodular set cover (ISSC) algorithm proposed in Iyer and Bilmes (2013), which offers $\frac{nH_{acc}}{1+(n-1)(1-\kappa_{unf})}$ bound on unf, where κ_{unf} is the curvature of unf, n is the number of features and H_{acc} is the approximation factor of the submodular set cover using the function acc (details on the bounds can be found in Iyer and Bilmes (2013)). As pointed out before, due to weak submodularity and non-monotonicity of acc, these bounds need not hold, but the empirical results obtained in the next section are very close to the optimum.

[6] Examples of cases where, for outcome fairness, one is legally bound to ensure that a decision making process yields the fairest solution under certain performance constraints are discussed at length in (Barocas and Selbst 2016).

5 Evaluation

Here, we evaluate the effectiveness of applying the constrained submodular optimization methods of Iyer and Bilmes (2013) to problems (8) and (9). An open-source code implementation is available at: http://fate-computing.mpi-sws.org/. We show that empirically these methods rapidly provide near-optimal tradeoffs between process fairness and accuracy.

5.1 Experimental setup & performance measures

We address problems (8) and (9) using the iterated submodular-cost knapsack (ISK) and iterated submodular set cover (ISSC) methods proposed by Iyer and Bilmes (2013). These methods require training classifiers on various feature subsets of the training data. To ensure (weak) submodularity of accuracy, we use logistic regression classifiers with L2-regularization (Elenberg et al. 2016).

To examine a broad range of tradeoffs between process fairness and accuracy, we obtain solutions for (8) and (9) using multiple thresholds for unfairness and accuracy respectively. For each problem, we use 21 different threshold values, covering the full range of possible values of accuracy and unfairness, with constant step size.

In each case, we compare the performance of the constrained submodular optimization method with the true optimum achieved by brute force exhaustive enumeration over all the 2^k possible classifiers, where k is the total number of features that are available in the dataset under consideration. As stated in Section 4, for the NYPD SQF dataset, we calculated the optimal results using brute force methods only for a subset of 16 most informative features. However, the scalability of the constrained submodular optimization methods allows us to approximate solutions for the full set of 30 features as well.

5.2 Results

We discuss results for maximizing accuracy subject to maximum feature-apriori unfairness constraints (8), as shown in Figure 4. The y-axis shows accuracy attained for approximate submodular and optimal (exhaustive search) methods, as the unfairness threshold is varied on the x-axis.Results for other notions of process fairness defined in Section 2, as well as the results for minimizing unfairness subject to accuracy (9) are qualitatively similar.

Observations. Our fast methods for constrained submodular optimization work very well empirically, achieving results that are close to optimal. This is encouraging since these methods are highly scalable. Below, we describe our observations in detail for the ProPublica COMPAS dataset. Similar holds for NYPD SQF data.

We observe several stages of the accuracy / fairness tradeoff in Figure 4 [Left] as the maximum unfairness threshold is varied. For an unfairness threshold of 0 (a perfectly fair classifier), an empty feature set is selected, which achieves the accuracy of the null classifier. As the unfairness threshold rises, accuracy first increases sharply, due to the addition of highly fair and informative features (“number of prior offenses”), and then slowly, by adding highly fair but less in-

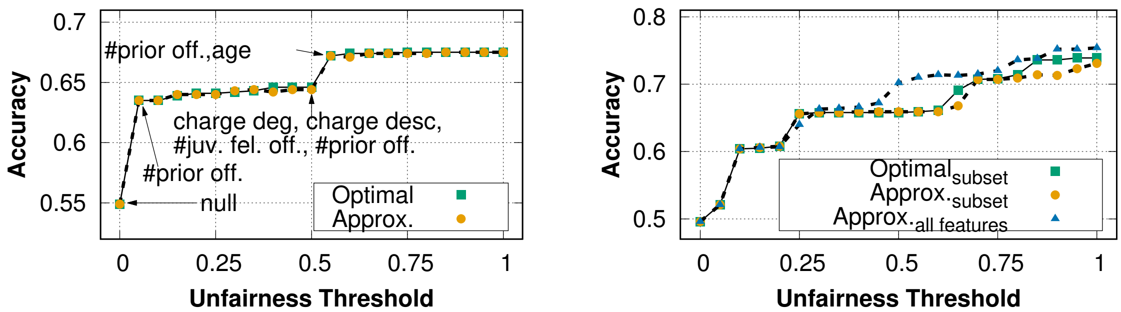

Figure 4: Finding a subset of features that maximizes accuracy, subject to unfairness constraints for [**Left**] the ProPublica COMPAS dataset and [**Right**] the NYPD SQF dataset. For a given maximum threshold of unfairness (x-axis), the plots show the accuracy (y-axis) of the classifier trained on the respective subset of features. The optimal (brute force) solutions for the 9 ProPublica COMPAS features and 16 most informative NYPD SQF features are shown in green, and the approximate solutions (Problem (8)) in yellow. Additionally, we show the approximate solutions for all 30 features from the NYPD SQF dataset in blue. In the left plot, the annotations with arrows show some specific features selected at various unfairness thresholds.

formative features (*e.g.*, "arrest charge description"). By increasing the threshold to 0.55, previously selected features are discarded, and a significantly unfair yet highly informative feature ("age") is added, leading to another sharp rise in accuracy. As the threshold rises higher, accuracy plateaus, since the remaining features do not add substantial predictive power.

We also note that when maximizing accuracy under fairness constraint, as the maximum unfairness constraint is relaxed, we do not simply see more features being gradually added and never removed. Rather, we sometimes see a significant change in the whole feature set used, in order to optimize the objective. Some features, such as "age", can have very high predictive power yet very low fairness.

Discussion. We comment on feasibility of the returned solutions. When we maximized accuracy subject to an unfairness constraint (8), for all runs, the fast method always returned a feasible solution (satisfying the constraint). However, when we minimized unfairness subject to a minimum accuracy threshold, occasionally the fast method returned solutions which were slightly infeasible, *i.e.*, the accuracy of the returned solution was marginally below the given threshold. For the ProPublica COMPAS dataset, this happened for $1/21$ runs; the maximum extent of infeasibility was 0.001 (*e.g.*, if a specified accuracy threshold of 0.65, 0.649 was returned). For the NYPD SQF dataset with 16 features, this happened $2/21$ times with maximum infeasibility of 0.001. For the full NYPD SQF dataset with 30 features, this happened $6/21$ times with maximum infeasibility of 0.006.

These empirical results suggest that it may be wise in practice to prefer the approach for maximizing accuracy subject to unfairness constraint, particularly for datasets with many features. However, we advise caution since in fact, if both unfairness and accuracy were exactly monotone and submodular, then the ISK algorithm for maximizing accuracy is not theoretically guaranteed to return a feasible solution (though infeasibility would be bounded), whereas when minimizing unfairness subject to accuracy with the ISSC algorithm, it is guaranteed to be feasible (Iyer and Bilmes 2013). Finally, note that if an infeasible solution is returned in practice, then it is simple just to move the threshold slightly and try again.

6 Process versus outcome fairness

Thus far, our evaluations have focused on our new process fairness measures, ignoring earlier measures of outcome fairness (Dwork et al. 2012; Feldman et al. 2015; Kamiran and Calders 2010; Luong, Ruggieri, and Turini 2011; Pedreschi, Ruggieri, and Turini 2008; Zafar et al. 2017b; 2017a; Zemel et al. 2013). We now examine empirically the relationship between process fairness and an established measure of outcome fairness, and consider their joint tradeoffs with accuracy.

The outcome fairness measure we use examines false positive and false negative rates for whites (w) and non-whites (nw). Specifically, we define:

$$\begin{aligned} &\text{outcome unfairness} = \\ &|FPR_w - FPR_{nw}| + |FNR_w - FNR_{nw}|, \\ &\text{outcome fairness} = \ -\text{ outcome unfairness}. \end{aligned} \tag{10}$$

Outcome fairness values can vary between -2 (very unfair) and 0 (very fair). This measure is inspired by recent studies related to fairness in criminal risk assessment (Kleinberg, Mullainathan, and Raghavan 2017; Angwin et al. 2016; Zafar et al. 2017a). Other measures of outcome fairness (*e.g.*, disparate impact considered by Zafar et al. (2017b)) could be used, but for the types of risk assessment analysis we consider, our definition may be more suitable (Angwin et al. 2016; Zafar et al. 2017a).

To study the tradeoff between process fairness, outcome fairness and accuracy, we train classifiers (optimizing for accuracy) with all possible combinations of features from the ProPublica COMPAS dataset ($2^9 = 512$ classifiers), as well as for a subset of 16 features from the NYPD SQF dataset ($2^{16} = 65,536$), and compute these three statistics (accuracy, process and outcome fairness) for all the classifiers.

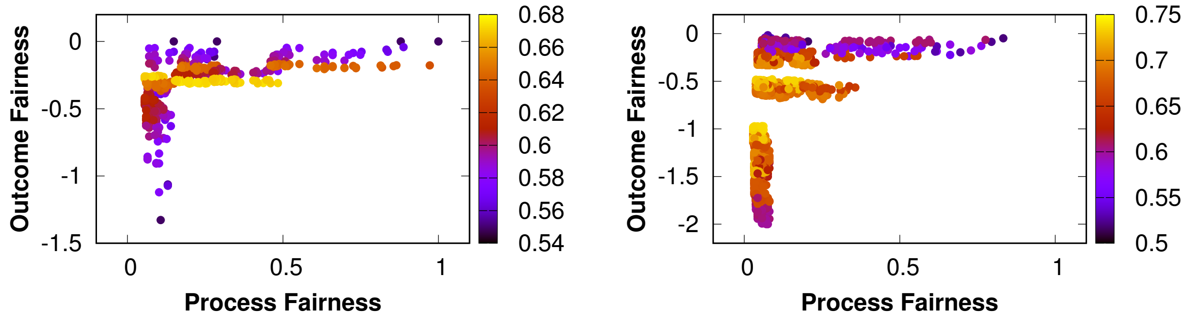

Figure 5: Tradeoffs between outcome fairness and feature-apriori fairness. The plots show the feature-apriori fairness (x-axis) and outcome fairness (y-axis), measured as disparity in mistreatment (Zafar et al. 2017a). The color of each point indicates the accuracy of the corresponding classifier. We report values for [**Left**] all $2^9 = 512$ ProPublica COMPAS classifiers, [**Right**] a random sample of 5000 out of $2^{16} = 65,536$ NYPD SQF classifiers. In these datasets, high process fairness appears to imply high outcome fairness.

Figure 5 shows the outcome fairness and accuracy vs. feature-apriori fairness (results for feature-accuracy and feature-disparity fairness are similar, omitted due to size constraints). Further to the right on the x-axis indicates greater process fairness, and further up on the y-axis indicates higher outcome fairness. The color of each point indicates the accuracy of the corresponding classifier. We make the following observations:

High process fairness appears to imply high outcome fairness. The plots in Figure 5 show a distinctive inverted L-shape: we observe that ensuring high process fairness leads to high outcome fairness. We emphasize that this may not hold for all datasets but it is striking that it holds so clearly here on both datasets.

Intuitively, process fairness seems to be a stronger notion, in that it imposes very restrictive constraints on the classifier. Examining our datasets, we found that the requirement of high process fairness is restricting the feature set to features which many people feel are fair. In our datasets, these features exhibited very low correlation (measured as mutual information) to race or other very sensitive features, and high process fairness thereby indirectly induced high outcome fairness. However, when features considered fair are correlated with sensitive features, high process fairness does not need to lead to outcome fairness.

Tradeoff with accuracy. As just observed, for our datasets, ensuring high process fairness always leads to high outcome fairness. Hence, it might be sufficient to optimize only the process fairness / accuracy tradeoff (using mechanisms discussed in Section 4) with at least moderately high process fairness threshold, which will thereby also reap high outcome fairness as a by-product.

In contrast, mechanisms proposed in prior fair learning works that optimize for outcome fairness and accuracy do not similarly guarantee high process fairness. Also, while these prior works have shown that one can maintain very high classification accuracy (close to the most accurate classifier) and still achieve a very high value of outcome fairness, the same is not true for process fairness. That is, the mechanisms for our datasets, and we imagine for many others, requiring high process fairness will prevent high classification accuracy, since this requirement will force informative features to be dropped (as explained in Section 5).

7 Conclusion

Much of the recent work on fairness in machine learning has either focused on analyzing outcomes, or has been inspired by and restricted to notions explored in the anti-discrimination literature, where specific features are deemed to be either sensitive (unusable) or not sensitive (usable). In this work, we complement earlier work by adding fairness notions beyond binary discrimination. Specifically, we introduce three new scalar measures of fairness that explicitly account for individuals' moral sense for whether or not it is fair to use various input features in the decision making process. We show how we can operationalize these notions by gathering human judgments about using features in the context of two real-world scenarios: recidivism risk estimation and prediction of illegal weapon possession. We show how finding a good tradeoff between process fairness and accuracy of a classifier can be modeled as fast, scalable, constrained submodular optimization problems over the set of features, and demonstrate good empirical performance. For the datasets we consider, our results show that when we optimize for high process fairness, we also achieve high outcome fairness as a byproduct. We do not expect this relationship to hold in general. In future work we aim to develop a deeper understanding of this phenomenon.

References

Agan, A. Y., and Starr, S. B. 2016. Ban the Box, Criminal Records, and Statistical Discrimination: A Field Experiment. In *University of Michigan Law & Economics Research Paper No. 16012*.

Ahmed, F.; Dickerson, J. P.; and Fuge, M. 2017. Diverse Weighted Bipartite b-Matching. *arXiv:1702.07134*.

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

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Ashkan, A.; Kveton, B.; Berkovsky, S.; and Wen, Z. 2015. Optimal Greedy Diversity for Recommendation. In *IJCAI*.

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

Beahrs, J. O. 1991. Volition, Deception, and the Evolution of Justice. *Bulletin of the American Academy of Psychiatry & the Law*.

Blackburn, S. 2003. *Being Good: A Short Introduction to Ethics*. OUP Oxford.

Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*.

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

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A Computational Approach to Politeness with Application to Social Factors. In *ACL*.

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

Elenberg, E. R.; Khanna, R.; Dimakis, A. G.; and Negahban, S. 2016. Restricted Strong Convexity Implies Weak Submodularity. *arXiv:1612.00804*.

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

Flores, A. W.; Lowenkamp, C. T.; and Bechtel, K. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.".

GDPR. 2016. GDPR Portal: Site Overview. http://www.eugdpr.org/.

Goel, S.; Rao, J. M.; and Shroff, R. 2015. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics*.

Graham, J.; Haidt, J.; Koleva, S.; Motyl, M.; Iyer, R.; Wojcik, S.; and Ditto, P. 2012. Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. *Advances in Experimental Social Psychology*.

Greenberg, J. 1987. A Taxonomy of Organizational Justice Theories. *Academy of Management Review*.

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

Iyer, R., and Bilmes, J. 2012. Algorithms for Approximate Minimization of the Difference Between Submodular Functions, With Applications. In *UAI*.

Iyer, R., and Bilmes, J. A. 2013. Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints. In *NIPS*.

Kamiran, F., and Calders, T. 2010. Classification with No Discrimination by Preferential Sampling. In *BENELEARN*.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware Classifier with Prejudice Remover Regularizer. *Machine Learning and Knowledge Discovery in Databases*.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding Discrimination through Causal Reasoning. In *NIPS*.

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

Krause, A.; Singh, A.; and Guestrin, C. 2008. Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *Journal of Machine Learning Research*.

Kusner, M. J.; Loftus, J. R.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *NIPS*.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. Data and Analysis for 'How We Analyzed the COMPAS Recidivism Algorithm'. https://github.com/propublica/compas-analysis.

Lin, H., and Bilmes, J. 2009. How to Select a Good Training-Data Subset for Transcription: Submodular Active Selection for Sequences. In *Interspeech*.

Lin, H., and Bilmes, J. 2011. A Class of Submodular Functions for Document Summarization. In *ACL HLT*.

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

Mason, W., and Suri, S. 2012. Conducting Behavioral Research on Amazon's Mechanical Turk. *Behavior Research Methods*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-Aware Data Mining. In *KDD*.

SQF Dataset. 2017. http://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page.

The Mechanical Turk Blog. 2011. Get Better Results with Less Effort with Mechanical Turk Masters. http://mechanicalturk.typepad.com/blog/2011/06/get-better-results-with-less-effort-with-mechanical-turk-masters-.html.

Tibshirani, R. 1994. Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society: Series B*.

Trankell, A. 1972. *Reliability of Evidence: Methods for Analyzing and Assessing Witness Statements*. Beckmans.

Yaari, M. E., and Bar-Hillel, M. 1984. On Dividing Justly. *Social Choice and Welfare*.

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

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

Zafar, M. B.; Gummadi, K. P.; and Danescu-Niculescu-Mizil, C. 2016. Message Impartiality in Social Media Discussions. In *ICWSM*.

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