The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making

Nina Grgić-Hlača
MPI-SWS, Germany
nghlaca@mpi-sws.org

Muhammad Bilal Zafar
MPI-SWS, Germany
mzafar@mpi-sws.org

Krishna P. Gummadi
MPI-SWS, Germany
gummadi@mpi-sws.org

Adrian Weller
University of Cambridge, UK
adrian.weller@eng.cam.ac.uk

Abstract

Machine learning methods are increasingly being used to inform, or sometimes even directly to make, important decisions about humans. A number of recent works have focussed on the *fairness* of the *outcomes* of such decisions, particularly on avoiding decisions that affect users of different sensitive groups (*e.g.*, race, gender) disparately. In this paper, we propose to consider the fairness of the *process* of decision making. Process fairness can be measured by estimating the degree to which people consider various features to be fair to use when making an important legal decision. We examine the task of predicting whether or not a prisoner is likely to commit a crime again once released by analyzing the dataset considered by ProPublica relating to the COMPAS system. We introduce new measures of people's discomfort with using various features, show how these measures can be estimated, and consider the effect of removing the uncomfortable features on prediction accuracy and on outcome fairness. Our empirical analysis suggests that process fairness may be achieved with little cost to outcome fairness, but that some loss of accuracy is unavoidable.

1 Motivation and New Measures of Process Fairness

As machine learning methods are increasingly being used in decision making scenarios that affect human lives, there is a growing concern about the *fairness* of such decision making. These concerns have spawned a flurry of recent research activity into learning methods for *detecting* and *avoiding* unfairness in decision making [8, 9, 12, 15, 16, 18, 19]. Our focus in this paper is on the foundational notions of *fairness* that underlie these fair learning and unfairness detection methods.

We argue that the notions of fairness underlying much of the prior work are centered around the **outcomes** of the decision process. They are inspired in large part by the application of anti-discrimination laws in various countries [2], under which decision policies or practices (implemented by humans) can be declared as discriminatory based on their effects on people belonging to certain sensitive demographic groups (*e.g.*, gender, race). For instance, the notions of "individual fairness" in [8], "situational testing" in [15], and "disparate treatment" in [18], consider individuals who belong to different sensitive groups, yet share similar non-sensitive features (qualifications), and require them to receive *similar decision outcomes*. Similarly, the notions of "group fairness" in [19] and "disparate impact" in [18] are based on different sensitive groups (*e.g.*, males and females or blacks and whites) receiving *beneficial decision outcomes in similar proportions*. Finally, the notion of "disparate mistreatment" in [17] is rooted in the desire for different sensitive demographic groups to

Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

experience *similar rates of errors in decision outcomes*. Thus, in prior works, the fairness of decision making has been evaluated based on the decision outcomes.

In this paper, we make the case for notions of fairness that are based on the **process** of decision making rather than on the outcomes. Our notions of process fairness are motivated by the observation that in many decision making scenarios, humans tend to have a *moral sense* for whether or not it is fair to use an input feature in the decision making process. For instance, consider the task of predicting recidivism risk for an offender. COMPAS is a commercial recidivism prediction tool that relies on a number of different types of user features, such as information about *Criminal history*, *Family criminality*, *Work* and *Social environment* of the offender. In a user survey that we conducted, we found that a strong majority of users felt that it was fair to use *Criminal history*, but unfair to use *Family criminality*. On the other hand, features *Work* and *Social environment* were deemed as fair and unfair (respectively) by only a weak majority of users.

Such societal consensus (strong or weak) on the fairness of using a feature in a decision process may be rooted in prevailing cultural or social norms, or political beliefs or legal (privacy) regulations or historical precedents. Unfortunately, existing outcome-based fairness notions developed for learning systems fail to capture this intuitive human understanding of fairness. Instead, current fair learning mechanisms justify the means (process) by the ends (outcomes), ignoring the different levels of societal consensus on the desirability of using different features in decision making. We propose different notions of process fairness to redress this situation.

1.1 Defining process fairness

Suppose a learning method, say a classifier $\mathcal{C}$, has been trained to make decisions using a set of features $\mathcal{F}$. Intuitively, the classifier's decision process would be considered fair by a user u *only* if the user u judges the use of every one of the features in the set $\mathcal{F}$ to be fair. We leverage this intuition to define the process fairness of the classifier $\mathcal{C}$ to be the fraction of all users that consider the use of every one of the features in $\mathcal{F}$ to be fair.

Our process fairness definition relies critically on users' judgments about the use of individual features when making decisions. Note that a user's judgment about a feature may change after they learn how using the feature might affect the decision outcomes. For instance, a user who initially considered a feature unfair for use in 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. Similarly, learning that using a feature might increase or decrease disparity in decision outcomes for different demographic groups (e.g., whites vs. blacks or men vs. women) might make a user change their opinion on the fairness of using that feature in decision making.

To capture the above concepts, where we recognize that users' judgments about using individual features in decision making might vary after they learn about impacts, we define three measures of process fairness: **feature-apriori fairness**, **feature-accuracy fairness** and **feature-disparity fairness**.

Consider a scenario for making some important decision. Let $\mathcal{U}$ denote the set of all members ('users') of society, and $\mathcal{F}$ denote the set of all possible features that might be used in the decision making process.

Feature-apriori fairness. For a given feature $f \in \mathcal{F}$, let $\mathcal{U}_f \subseteq \mathcal{U}$ denote the set of all users that consider the feature f fair to use without *a priori* knowledge of how its usage affects outcomes. Given a set of features $\mathcal{F}'$, let $\mathcal{C}_{\mathcal{F}'}$ denote the classifier that uses those features $\mathcal{F}'$. We define

$$\text{feature-apriori fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{|\bigcap_{f_i \in \mathcal{F}'} \mathcal{U}_{f_i}|}{|\mathcal{U}|}. \tag{1}$$

Feature-accuracy fairness. Let $\mathcal{U}_f^{Acc} \subseteq \mathcal{U}$ denote the set of all users that consider the feature f fair to use *if it increases the accuracy of the classifier*. Note that typically we expect $\mathcal{U}_f \subseteq \mathcal{U}_f^{Acc}$, though this need not always hold exactly (either due to noise in estimating user preferences, or due to some users attaching some sort of negative connotation to the notion of accuracy).[1] Given a set of features $\mathcal{F}'$, let $\mathcal{C}_{\mathcal{F}'}$ denote the classifier that uses those features $\mathcal{F}'$, and let $Acc(\mathcal{C}_{\mathcal{F}'})$ denote its accuracy. We define

$$\text{feature-accuracy fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{|\bigcap_{f_i \in \mathcal{F}'} Condition(\mathcal{U}_{f_i}, \mathcal{U}^{Acc}_{f_i})|}{|\mathcal{U}|}, \tag{2}$$

where

$$Condition(\mathcal{U}_{f_i}, \mathcal{U}^{Acc}_{f_i}) = \begin{cases} \mathcal{U}_{f_i} \cup \mathcal{U}^{Acc}_{f_i}, & \text{if } Acc(\mathcal{C}_{\mathcal{F}'}) > Acc(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}) \\ \mathcal{U}_{f_i}, & \text{if } Acc(\mathcal{C}_{\mathcal{F}'}) \leq Acc(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}). \end{cases} \tag{3}$$

Feature-disparity fairness. Let $\mathcal{U}^{Disp}_f \subseteq \mathcal{U}$ denote the set of all users that consider the feature f fair to use *even* if it increases a measure of disparity (i.e. disparate impact or disparate mistreatment) of the classifier. Typically we expect $\mathcal{U}^{Disp}_f \subseteq \mathcal{U}_f$, though this need not always hold strictly due to estimation error or other reasons.[2] Given a set of features $\mathcal{F}'$, let $\mathcal{C}_{\mathcal{F}'}$ denote the classifier that uses those features $\mathcal{F}'$, and let $Disp(\mathcal{C}_{\mathcal{F}'})$ denote the disparity it induces. We define

$$\text{feature-disparity fairness of } \mathcal{C}_{\mathcal{F}'} := \frac{|\bigcap_{f_i \in \mathcal{F}'} Condition(\mathcal{U}_{f_i}, \mathcal{U}^{Disp}_{f_i})|}{|\mathcal{U}|}, \tag{4}$$

where

$$Condition(\mathcal{U}_{f_i}, \mathcal{U}^{Disp}_{f_i}) = \begin{cases} \mathcal{U}_{f_i} \cup \mathcal{U}^{Disp}_{f_i}, & \text{if } Disp(\mathcal{C}_{\mathcal{F}'}) \leq Disp(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}) \\ \mathcal{U}^{Disp}_{f_i}, & \text{if } Disp(\mathcal{C}_{\mathcal{F}'}) > Disp(\mathcal{C}_{\mathcal{F}' \setminus \{f_i\}}). \end{cases} \tag{5}$$

1.2 Contributions

We highlight the main contributions of this paper.

- We define three different measures of process fairness.
- We show how the definitions can be operationalized in the context of recidivism risk estimation using the public COMPAS ProPublica dataset [14]. We assemble a ranked list of user preferences for features, which allows us to explore measures of process fairness.
- We provide an initial analysis of the trade-offs between accuracy and different measures of fairness when using different subsets of these features. We develop several avenues for future work.

1.3 Related work

Earlier work on fairness [18] focused primarily on achieving high decision-making accuracy without using sensitive features (*e.g.*, gender, race) which are legally prohibited from being used in the decision-making process [2]. However, in this paper, we present a broader notion of fairness that goes beyond the current legal specifications by examining the extent to which people feel that each feature (sensitive or non-sensitive) is unfair to use and how removing a combination of these features would affect the outcome fairness and accuracy.

We now briefly comment on how our work is inspired by and is related to works in other disciplines such as social sciences, moral philosophy, political sciences, and law. In moral philosophy [6]: 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 our notion of outcome fairness.

[1] The set $\mathcal{U}_f \setminus \mathcal{U}^{Acc}_f$ consists of those peculiar users who consider a feature fair to use *only if it does not increase the accuracy of the classifier*. As such, these users might be considered outliers that should be excluded (though we include them here). Across all our survey questions, less than 5% of users were peculiar in this way.

[2] Similarly to footnote 1, the set $\mathcal{U}^{Disp}_f \setminus \mathcal{U}_f$ consists of peculiar users (included in our analysis here) who consider a feature fair to use *only if it increases a measure of disparity of the classifier*. For all our survey questions, less than 5% of users were peculiar in this way.

Prior literature in social, economic, legal, and political sciences distinguishing between direct discrimination and indirect discrimination makes similar observations as we do in this paper. These works point out that the "wrong" of direct (process) discrimination should be distinguished from the "wrong" of indirect (outcome) discrimination [3]. Similarly, we argue that when considering decision-making, fairness of the process is distinct from fairness of the outcome.

2 Measuring process fairness

In this section, we use the ProPublica COMPAS dataset [14] and construct a classification task where the goal is to predict whether a criminal defendant would commit a crime again (recidivate) or not. We chose to focus on this specific dataset since it is publicly available, and its findings related to possible racial bias led to a considerable number of follow up studies [7, 10, 13, 17].

We measure the process fairness (using the definitions described in Section 1.1) for all of the possible classifiers that can be constructed from different combinations of features present in the dataset. Since the measurement of process fairness relies on human judgments, we first gather human judgments on the fairness of individual features as described in Section 1.1. Then, we use these judgments to quantify process fairness for all possible classifiers.

For building classifiers corresponding to each combination of features, we use logistic regression. To obtain reliable estimates of classifier performance, we split the data randomly into train (50%) and test (50%) folds 10 times and report the averages (results for each of the 10 runs were similar).

2.1 Gathering human judgments of individual feature fairness

Dataset. We examine the ProPublica COMPAS dataset, which consists of all criminal defendants who were subject to COMPAS screening in Broward County, Florida, during 2013 and 2014. For each defendant, various information fields ('features') were also gathered by ProPublica. Broadly, these fields are related to the defendant's demographic information (*e.g.*, gender and race), criminal history (*e.g.*, the number of prior offenses) and administrative information about the case (*e.g.*, the case number, arrest date, risk of recidivism predicted by the COMPAS tool). Finally, the dataset also contains information about whether the defendant did actually recidivate or not.[3]

For this analysis, we are only interested in features that could potentially be used to predict the risk of recidivism of a defendant. Hence, we ignore the administrative features and use a subset of features for our analysis. As a result, we get 9 features that can be used to construct a classifier for recidivism prediction. These features are: arrest charge description (*e.g.*, grand theft, possession of drugs), charge degree (misdemeanor or felony), number of prior criminal offenses, number of juvenile felony offenses, juvenile misdemeanor offenses, other juvenile offenses, age of the defendant, gender of the defendant and race of the defendant.

Next, we describe the survey setup used to gather human judgments on the fairness of these features.

Survey setup. We recruited 100 Amazon Mechanical Turk (AMT) workers. Since the task relates 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 [4].[4]

Each AMT worker was shown the nine features described above, and for each feature, they were asked the following questions (see https://people.mpi-sws.org/~nghlaca/MLLaw_2016/process_fairness/):

> **Q. 1:** Do you believe it is fair or unfair to use information about this feature when estimating the offender's risk of recidivism?

[3] Aside from these added fields, the ProPublica features are a subset of the significantly larger feature set used by the COMPAS risk assessment tool for recidivism prediction [1]. The ProPublica dataset is publicly available for analysis, but the COMPAS dataset is not publicly available.

[4] It is possible that these AMT workers might not constitute a representative sample of the population. However, estimating representative demographics (e.g. various age groups, groups with different educational levels) and collecting matching data can be a challenging task, beyond the the scope of this work.

Feature	**Q. 1** (a priori)	**Q. 2** (if more accurate)	**Q. 3** (if increases disparity)
# prior offenses	95%	93%	83%
arrest charge description	86%	92%	71%
arrest charge degree	85%	91%	69%
# juvenile felony offenses	74%	80%	61%
# juvenile misdemeanor offenses	65%	71%	53%
# juvenile other offenses	63%	69%	52%
age	44%	61%	32%
gender	26%	55%	24%
race	21%	42%	17%

Table 1: Comparing user judgment of fairness of each feature, when the user has different knowledge about the impact of incorporating that feature in the decision making process. We show the percentage of users who categorized each feature as fair according to the 3 questions described in Section 2.1.

Q. 2: Do you believe it is fair or unfair to use information about this feature when estimating the offender's risk of recidivism, *if it makes the estimation more accurate*?

Q. 3: Do you believe it is fair or unfair to use information about this feature when estimating the offender's risk of recidivism, *if it makes black people more likely to be assessed as having a higher risk of recidivism than white people*?

Note that the answers to these questions for a given feature will be used to measure feature-apriori, feature-utility and feature-disparity measures of fairness, respectively, for classifiers using various subsets of the features.

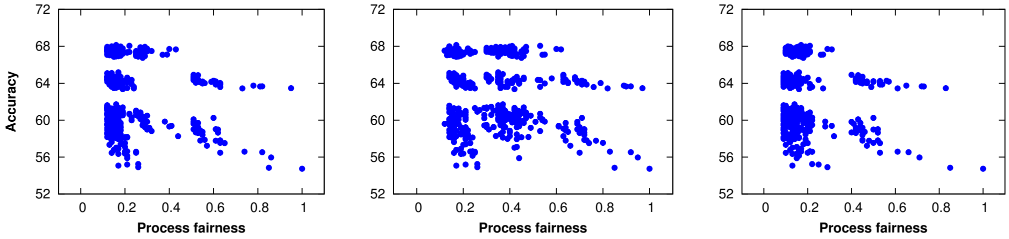
Classifier	**Acc.**	**f.-apriori f.**	**f.-acc. f.**	**f.-disp. f.**
Classifier with no features	55.56%	1.00	1.00	1.00
Most accurate classifier	68.06%	0.12	0.34	0.1
Most fair classifier	63.04%	0.95	0.97	0.84

Table 2: Accuracy and process fairness statistics for a null classifier, most accurate classifier (out of 512 classifiers) and most fair classifier with respect to all three measures of process fairness. The measures of process fairness are feature-apriori fairness (f.-apriori f.), feature-accuracy fairness (f.-acc. f.) and feature-disparity fairness (f.-disp. f.).

Comparing human judgments of fairness. For each of the 9 features, we computed the fraction of AMT workers who judged that feature to be fair under each of the knowledge settings described in questions Q. 1, 2 and 3. The overall results are shown in Table 1. In addition, we divided the respondents into subgroups based on various categories (white/non-white and male/female), but for each subgroup the results were similar. See the Appendix for details. We make the following observations on the results.

First, when asked about the a priori fairness of a feature (Q. 1), the fraction of AMT workers who judged each feature to be fair varied significantly across features. Features related to the criminal history and the current crime were judged to be fair by almost all workers. Features related to juvenile offenses were deemed fair by fewer workers, though still a majority. Finally, the features of age, gender and race, which might sometimes be considered as protected [2] were deemed highly unfair (a majority of the workers judged them to be unfair) with the 'race' feature considered fair by just 21% of the workers. Thus, the features which are judged most morally unfair are also the ones that are sometimes protected by the law [2]. However, our results indicate a more nuanced, scalar view of the judged fairness of features.

Second, regarding the judged fairness of a feature when it is known that its use would lead to an increase in prediction accuracy (Q. 2), as expected, the judged fairness of the feature increases as compared to Q. 1 (this holds for all features except for the number of prior offenses, where the difference is barely significant). Specifically, the percentage of workers who judged the 'race' feature



(a) Feature-apriori fairness (b) Feature-accuracy fairness (c) Feature-disparity fairness

Figure 1: Accuracy vs. different measures of process fairness.

to be fair when they know that it increases accuracy doubles (42%) as compared to when no additional knowledge is provided (21%).

Third, the judged fairness of a feature when it is revealed that its usage would increase the racial disparity of outcomes, is significantly lower than in the first two cases, confirming expectations that a potential for disparity lowers human judgment of fairness.

Finally, we observe that the relative ranking of the features remains constant for all knowledge settings (Q. 1, 2 and 3).

2.2 Leveraging human judgments to measure process fairness

Using the human judgments described above, we computed the three measures of process fairness defined in Section 1.1—feature-apriori fairness, feature-accuracy fairness and feature-disparity fairness—for classifiers that can be constructed using all possible different combinations of features.

Specifically, since the dataset has 9 features, there are $2^9 = 512$ different classifiers that can be trained using all possible subsets of the features. For each of the classifiers, we also compute the empirical accuracy and outcome fairness on the test set. Our notion of outcome fairness is defined in Section 3.2.

Next, we present the null classifier (which has no features), the classifier with the best accuracy and the most fair classifier with respect to all three measures of process fairness. For the selected classifiers, the accuracy and all three measures of process fairness are shown in Table 2.

The null classifier serves as a baseline, and in this case it achieves an accuracy of 56%.

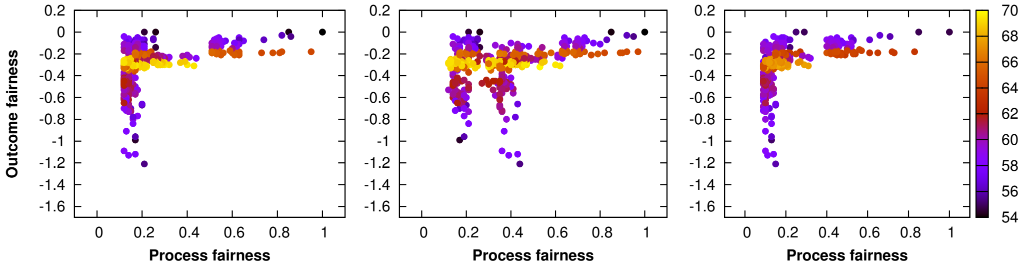
The most accurate classifier achieves an accuracy of 68.1%. However, it is very unfair with respect to all three types of process fairness.

The most fair classifier with respect to all three types of process fairness is the one that uses the feature ‘number of prior criminal offenses’. Even though this classifier uses only one feature, it has the highest feature fairness and a modest accuracy of 63.0%.

Let us also point out a few key observations explaining how using the feature ‘race’ affects the fairness of the classification process. The feature ‘race’ is deemed as a very unfair feature, ranking last in **Q. 1**, **2** and **3**. As such, it inflicts an upper bound on the maximum value of all three types of process fairness, 0.21, 0.42 and 0.17 respectively. However, even though the process fairness is low, the accuracy has a range from 54.59% to 66.93%.

3 The Cost of Process Fairness: Outcome Fairness and Accuracy

While removing an undesirable feature improves process fairness, it may also lead to reduced accuracy or lower outcome fairness. In this section, we conduct an initial empirical analysis of these tradeoffs, saving a more complete analysis for future work. Our approach is to train classifiers (optimizing to achieve highest prediction accuracy) using each of the $2^9 = 512$ possible choices of subsets of the 9 features, and examine their accuracy, process fairness and outcome fairness.



(a) Feature-apriori fairness (b) Feature-accuracy fairness (c) Feature-disparity fairness

Figure 2: Outcome fairness, measured as disparity in mistreatment, vs. different measures of process fairness for different classifiers. The color intensity of each point represents the accuracy of the corresponding classifier.

3.1 Process fairness and accuracy tradeoff

We present the accuracy of each of 512 classifiers in Figure 1 against the three notions of fairness: feature-apriori fairness, feature-accuracy fairness and feature-disparity fairness for our results. We make the following interesting observations:

First, for very low value of process fairness (vertically spread cluster on the left side in all three figures), one could achieve almost all possible values of accuracy. Upon further investigation, we find out that the these clusters correspond to the classifiers when features judged to be highly unfair (gender, race and age) are present in the classifier feature set.

On the other hand, the right end of the three figures (corresponding to the scenarios when the process fairness is high and larger than 0.90) correspond to the cases when the three features (gender, race and age) deemed to be highly unfair (Table 1) are not present in the corresponding classifiers' feature sets. The maximum accuracy achieved by such classifiers is 64.5%, which is somewhat lower than the accuracy of the best possible classifier (68.1%). Our findings suggest a tradeoff between process fairness and accuracy, where to achieve high process fairness, we need to drop certain features from classifiers, which in turn leads to a drop in accuracy. It is possible to achieve **high accuracy or high process fairness, but not both simultaneously.**

3.2 Process fairness, outcome fairness, and accuracy tradeoff

Next, we study the tradeoffs between process fairness and outcome fairness. Inspired by recent studies analyzing the ProPublica dataset [5, 17], we define our measure of outcome fairness in terms of disparities in the misclassification rates for Whites (w) and non-Whites (nw).[5] Specifically, we first define a measure of *outcome (un)fairness* as follows:

$$outcome\ unfairness = |FPR_w - FPR_{nw}| + |FNR_w - FNR_{nw}|. \tag{6}$$

$$outcome\ fairness = -outcome\ unfairness. \tag{7}$$

Thus, outcome fairness is a negative value, with higher values (closer to 0) indicating greater fairness, in the sense of lower discrepancy between respective false positive and false negative rates. Other measures of outcome fairness (*e.g.*, disparate impact) could be used but, as pointed out by recent studies, for the risk assessment analysis [5, 17], our disparity in misclassification rates might be a more suitable notion of outcome fairness for this dataset.

We computed the values of *outcome fairness* for each of the 512 classifiers and compared them to the corresponding values of process fairness. Figure 2 shows the outcome fairness vs. the three notions

[5] The dataset contains multiple races: Black (51%), White (34%), Hispanic (9%), other (5%), Asian (0.4%) and Native American (0.2%). However, since it is easier to define an outcome fairness measure for binary values, we convert the races to White and non-White. This categorization is in line with the ProPublica finding which claimed that Whites get preferential treatment compared to other races.

of process fairness defined in Section 1. Note that increasing values towards the right hand side on the x-axis indicate a higher level of process fairness, and increasing values towards the upwards side on the y-axis indicate a higher level of outcome fairness. Analyzing Figure 2, we make the following observations:

Comparing Figures 2(a), (b), and (c), corresponding to different notions of process fairness: we notice that the long vertical cluster on the left side of Figure 2(a) (low feature-apriori fairness) splits into two long vertical sub-clusters in Figure 2(b) (feature-accuracy fairness). These sub-clusters can be explained by the fact that some of the features that were judged unfair by AMT workers at first, were considered fair if they led to an increase in accuracy, hence increasing the feature-accuracy fairness. Similarly, we notice that Figures 2(a) and 2(c) are qualitatively very similar, except that Figure 2(c) presents relatively lower values of process fairness as compared to 2(a). This observation can be explained by feature-disparity fairness (judged with the knowledge that a feature usage would led to greater disparity in outcomes) being consistently lower than feature-apriori fairness (judged without the knowledge of how a feature's usage would impact disparity).

Next, we see that many points with high accuracy (represented by the intensity of their color) are clustered around the top left corner of the three plots. These points correspond to the classifiers that achieve high outcome fairness but low process fairness. Included in these points is the classifier that uses the features age, juvenile misdemeanor count, juvenile other count, priors count, race and sex, and achieves the best accuracy among all classifiers (68.1%) while achieving a high outcome fairness (0.23). The process fairness measures for this classifier are abysmal 0.12, 0.34 and 0.08, respectively. This observation shows that **achieving high accuracy and high outcome fairness simultaneously comes at a steep cost of low process fairness**.

The purple point in the extreme top-right corner of all the plots corresponds to the null classifier which uses no features and hence achieves perfect process and outcome fairness. However, this classifier is not informative in terms of predicting recidivism. The point below that in all figures achieves very high values of process and outcomes fairness, while retaining a moderate accuracy of 63.0% (as compared to the null classifier accuracy of 56% and the best possible classifier accuracy of 68.1%). This point corresponds to a classifier that only uses the feature 'number of prior criminal offenses'. This observation shows that **achieving high process fairness and high outcome fairness simultaneously comes at the cost (albeit moderate) of accuracy in predictive power**.

3.3 Jointly optimizing for process fairness, outcome fairness and accuracy

A number of recent works have proposed learning mechanisms to optimize jointly for both accuracy and outcome fairness (rather than just for accuracy). These works have reported promising results, where they achieve high outcome fairness at a relatively small cost on accuracy. It would be interesting to investigate learning methods that instead optimize jointly for accuracy and process fairness, or even for all three measures of accuracy, process fairness and outcome fairness. This optimization will involve computational challenges in addition to interesting statistical issues. We leave this analysis for future work but make a few observations. First, our empirical findings above suggest that achieving high accuracy (and high outcome fairness) and high process fairness at the same time might be much harder. Second, an attractive property of all our definitions of process fairness that might facilitate analysis and optimization, is that all our measures of process fairness are *submodular* [11].

4 Conclusion

While previous work has focused on exploring measures of *outcome* fairness, that is how to avoid disparate treatment, disparate impact, or disparate mistreatment, here we introduced three quantitative measures of *process* fairness. We showed how these measures can be obtained in an important legal application. In exploring the tradeoff between accuracy and fairness, an exciting conclusion of earlier work was that it appears that there are practical situations where outcome fairness can be achieved with only very slight cost to accuracy. In contrast, our empirical analysis of the ProPublica COMPAS dataset suggests that in order to achieve process fairness, a more significant amount of accuracy must be lost (though there is good news that higher process fairness has little cost in terms of outcome fairness). We plan to explore the tradeoffs between process fairness, outcome fairness and accuracy further in future work.

References

[1] https://www.propublica.org/documents/item/2702103-Sample-Risk-Assessment-COMPAS-CORE.

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

[3] Discrimination — Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/discrimination/, 2011.

[4] Get better results with less effort with Mechanical Turk Masters – The Mechanical Turk blog. http://bit.ly/amt-masters, 2011.

[5] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. 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, 2016.

[6] S. Blackburn. *Being good: A short introduction to ethics*. OUP Oxford, 2003.

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

[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] A. W. Flores, C. T. Lowenkamp, and K. Bechtel. 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.". 2016.

[11] S. Fujishige. *Submodular functions and optimization*, volume 58. Elsevier, 2005.

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

[13] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[14] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. Data and analysis for 'How we analyzed the COMPAS recidivism algorithm'. https://github.com/propublica/compas-analysis, 2016.

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

[16] D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware Data Mining. In *KDD*, 2008.

[17] M. B. Zafar, I. V. Martinez, M. G. Rodriguez, and K. P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *arXiv:1610.08452*, 2016.

[18] M. B. Zafar, I. V. Martinez, M. G. Rodriguez, and K. P. Gummadi. Learning Fair Classifiers. *arXiv:1507.05259*, 2016.

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

A Additional Results

In table 1, we presented the results of our analysis of user judgments of individual feature fairness, conducted on 100 AMT workers, for 9 features from the dataset considered by ProPublica, relating to the COMPAS system. Now we compare user judgments of fairness of individual features, for different demographic subgroups of people.

We conduct two separate comparisons: one for men and women, and the other for white and non-white AMT workers. The results show that the rankings of features are quite stable across different subsets of the population.

Tables 3 and 4 show a general similarity between the judgments of 54 men and 46 women, as well as the similarity of the judgments of these subgroups to the majority vote ranking, for all three questions Q. 1, 2 and 3. However, there are some differences. In all three questions, fewer female than male workers have judged that the three features related to juvenile offenses are fair to use. Also, we can observe slight differences in the relative rankings of features, due to minor differences between the fraction of male and female workers that judged specific features to be fair.

Tables 5 and 6 show how 84 white and 16 non-white people rated the fairness of the same features. We are comparing white with non-white workers, aggregating all non-white races, since they constitute only a small fraction of the workers (e.g., 3 black workers). The results of the analysis show that the judgments of white workers coincide with the general trends observed in the other tables. On the other hand, the judgments of non-white workers have produced a relative ranking of features that is less consistent with the prior observations.

One such case is the feature ‘gender’ (which is usually considered to be a sensitive feature), which ranks quite high on the non-white workers’ list, as a fair feature to use if it increases accuracy (Q. 2). Another example is the pair of features ‘arrest charge description’ and ‘number of prior offenses’. In all three questions Q. 1, 2 and 3, non-white workers have ranked the feature ‘arrest charge description’ as more fair to use than the feature ‘number of prior offenses’. However, our dataset contains judgments of only 16 non-white workers, so these differences might be attributed to the small sample size.

Feature	**Q. 1**	**Q. 2**	**Q. 3**
	(a priori)	(if more accurate)	(if increases disparity)
# prior offenses	96%	94%	87%
Arrest charge description	83%	93%	72%
Arrest charge degree	85%	91%	74%
# juvenile felony offenses	80%	80%	69%
# juvenile misdemeanor offenses	72%	74%	61%
# juvenile other offenses	70%	74%	61%
Age	46%	56%	32%
Gender	24%	50%	22%
Race	24%	35%	15%

Table 3: Comparing judgment of fairness of each feature, of 54 **male** users, when they have different knowledge about the impact of incorporating that feature. We show the percentage of users who categorized each feature as fair according to the 3 questions described in §2.1.

Feature	**Q. 1** (a priori)	**Q. 2** (if more accurate)	**Q. 3** (if increases disparity)
# prior offenses	94%	91%	78%
Arrest charge description	89%	91%	70%
Arrest charge degree	85%	91%	63%
# juvenile felony offenses	67%	80%	52%
# juvenile misdemeanor offenses	57%	67%	44%
# juvenile other offenses	54%	63%	41%
Age	41%	67%	33%
Gender	28%	61%	26%
Race	17%	50%	20%

Table 4: Comparing judgment of fairness of each feature, of 46 **female** users, when they have different knowledge about the impact of incorporating that feature. We show the percentage of users who categorized each feature as fair according to the 3 questions described in §2.1.

Feature	**Q. 1** (a priori)	**Q. 2** (if more accurate)	**Q. 3** (if increases disparity)
# prior offenses	96%	95%	87%
Arrest charge description	85%	91%	71%
Arrest charge degree	86%	92%	70%
# juvenile felony offenses	75%	81%	62%
# juvenile misdemeanor offenses	67%	74%	55%
# juvenile other offenses	66%	71%	54%
Age	42%	63%	33%
Gender	24%	54%	23%
Race	18%	43%	18%

Table 5: Comparing judgment of fairness of each feature, of 84 **white** users, when they have different knowledge about the impact of incorporating that feature. We show the percentage of users who categorized each feature as fair according to the 3 questions described in §2.1.

Feature	**Q. 1** (a priori)	**Q. 2** (if more accurate)	**Q. 3** (if increases disparity)
# prior offenses	88%	81%	63%
Arrest charge description	93%	100%	69%
Arrest charge degree	81%	88%	63%
# juvenile felony offenses	69%	75%	56%
# juvenile misdemeanor offenses	56%	56%	44%
# juvenile other offenses	50%	56%	44%
Age	56%	50%	25%
Gender	38%	63%	31%
Race	38%	38%	13%

Table 6: Comparing judgment of fairness of each feature, of 16 **non-white** users, when they have different knowledge about the impact of incorporating that feature. We show the percentage of users who categorized each feature as fair according to the 3 questions described in §2.1.