THE IMPACTS OF LABELING BIASES ON FAIRNESS CRITERIA

Yiqiao Liao
Department of Computer Science
The Ohio State University
liao.489@osu.edu

Parinaz Naghizadeh
Electrical and Computer Engineering
The Ohio State University
naghizadeh.1@osu.edu

ABSTRACT

As we increasingly rely on artificially intelligent algorithms to aid or automate decision making, we face the challenge of ensuring that these algorithms do not exhibit or amplify our existing social biases. An issue complicating the design of such fair AI is that algorithms are trained on datasets that can themselves be tainted due to the social biases of prior (human or AI) decision makers. In this paper, we investigate the robustness of existing (group) fairness criteria when an algorithm is trained on data that is biased due to errors by prior decision makers in identifying qualified individuals from a disadvantaged group. This can be viewed as labeling bias in the data. We first analytically show that some constraints such as Demographic Parity remain robust when facing such statistical biases, while others like Equalized Odds are violated if trained on biased data. We also analyze the sensitivity of the firm's utility to these biases under each constraint. Finally, we provide numerical experiments on three real-world datasets (the FICO, Adult, and German credit score datasets) supporting our analytical findings.

1 INTRODUCTION

Algorithmic decision making is being adopted widely in areas ranging from recommendation systems and ad-display, to hiring, loan approvals, and determining recidivism in courts. Despite their potential benefits, these algorithms can still exhibit or amplify existing social biases. There are at least two reasons for algorithmic bias: biases in the training datasets, and biases induced by the way the algorithm selects its decision rules (Mehrabi et al., 2021). More specifically, we use the term social bias or unfairness to mean that an algorithm makes decisions in favor or against individuals in a way that is inconsistent across groups with different social identities (e.g. race); a variety of fairness criteria (e.g., demographic parity, equalized odds) have been proposed with the goal of assessing and preventing these forms of unfairness. The term statistical bias on the other hand refers to biases in the datasets used to train the algorithm.

Our work explores the impacts of statistical biases on the efficacy of existing fairness criteria in addressing social biases. In particular, we investigate, both analytically and numerically, whether/which fairness criteria remain robust to biased training data generated by a prior decision maker who made errors when assessing qualified individuals from the disadvantaged group.

We first analytically show that some existing fairness criteria (Demographic Parity and True Positive Rate Parity) exhibit more robustness to such statistical biases compared to others (Equalized Odds and False Positive Rate Parity). We elaborate on the source of this difference by noting how prior qualification assessment biases impact the training data, and contrasting between the population characteristics that are used by each fairness criterion. We then artificially induce label biases in three real-world datasets (FICO, Adult, and German credit score), and numerically support our analytical findings. Our results can serve as an additional guideline when choosing among existing fairness criteria, or when proposing new criteria. They can inform decision makers who suspect that their available data may suffer from statistical biases, and who can accordingly select robust fairness criteria that would continue to be met in spite of any potential data biases.

Related work. Our work is related to those of (Ensign et al., 2018; Nie et al., 2018; Neel & Roth, 2018; Bechavod et al., 2019; Kilbertus et al., 2020; Wei, 2021; Yang et al., 2021; Blum & Stangl,

2020; Jiang & Nachum, 2020), who have investigated the interplay between data biases and fair algorithmic decision making. Most of these works differ from ours in that they explore how feedback loops, censored feedback, and/or adaptive data collection lead to statistical data biases, how these exacerbate algorithmic unfairness, and how to debias data. Our work is most similar to those of Blum & Stangl (2020) and Jiang & Nachum (2020), who also study the interplay between labeling biases and algorithmic fairness rules. However, the emphasis of these works is on proposing (reweighting) techniques to correct for label biases. In contrast, we assess the robustness of existing group fairness criteria, taking biases as fixed, with the interpretation that robust criteria are needed if statistical biases may go unnoticed, or while the debiasing efforts are in progress. Our findings are therefore complementary to these existing works.

2 Problem Setting

The agents. Consider a population of agents composed of two groups distinguished by a sensitive attribute $g \in \{a, b\}$ (e.g., race, gender). Let $n_g := \mathbb{P}(G = g)$ denote the fraction of group g agents. Each agent has an observable feature $x \in \mathbb{R}$, representing, e.g., exam scores or credit scores. (Our experiments consider both one-dimensional and n-dimensional features.) The agent further has a (hidden) binary qualification state $y \in \{0, 1\}$, with $y = 1$ and $y = 0$ denoting those that are qualified and unqualified to receive favorable decisions, respectively. Let $\alpha_g := \mathbb{P}(Y = 1|G = g)$ denote the qualification rate in group g. In addition, let $f_g^y(x) := \mathbb{P}(X = x|Y = y, G = g)$ denote the probability density function (pdf) of the distribution of features for individuals with qualification state y from group g. We make the following assumption on these feature distributions. In words, this assumption means that an individual is more likely to be qualified as its feature (score) increases.

Assumption 1. *The pdfs $f_g^y(x)$ and their CDFs $F_g^y(x)$ are continuously differentiable, and satisfy the strict monotone likelihood ratio property, i.e., $\frac{f_g^1(x)}{f_g^0(x)}$ is strictly increasing in $x \in \mathbb{R}$.*

We further define the qualification profile of group g as $\gamma_g(x) := \mathbb{P}(Y = 1|X = x, G = g)$, which captures the likelihood that an agent with feature x from group g is qualified. For instance, this could capture estimated repay probabilities given the observed credit scores (which may differ across groups). We will let group b be the group with a lower likelihood of being qualified at the same feature (i.e., if $\gamma_b(x) \leq \gamma_a(x), \forall x$), and refer to this group as the disadvantaged group.

The firm. A firm makes binary accept/reject decisions on agents based on their observable features. The firm gains a benefit of u_+ from accepting qualified individuals, and incurs a loss of u_- from accepting unqualified individuals. In this paper, we restrict attention to threshold policies $\pi_g(x) = 1(x \geq \theta_g)$, where $1(\cdot)$ denotes the indicator function and θ_g is the decision threshold for group g. Prior work (Liu et al., 2018; Zhang et al., 2020) show that this is without loss of generality under mild assumptions. Let $U(\theta_a, \theta_b) = n_a U_a(\theta_a) + n_b U_b(\theta_b)$ denote the firm's expected payoff.

The firm may further impose a (group) fairness constraint on the choice of its decision rule. While our framework is more generally applicable, we focus our analysis on `Demographic Parity (DP)` and `True/False Positive Rate Parity (TPR/FPR)`. Let $\mathcal{C}_a^{\mathtt{f}}(\theta_a) = \mathcal{C}_b^{\mathtt{f}}(\theta_b)$ denote the fairness constraint,[1] where $\mathtt{f} \in \{\mathtt{DP}, \mathtt{TPR}, \mathtt{FPR}\}$, and `DP` is given by $C_g^{\mathtt{DP}}(\theta) = \int_\theta^\infty (\alpha_g f_g^1(x) + (1 - \alpha_g) f_g^0(x))\mathrm{d}x$; `TPR`, also known as Equality of Opportunity (Hardt et al., 2016), is given by $\mathcal{C}_g^{\mathtt{TPR}}(\theta) = \int_\theta^\infty f_g^1(x)\mathrm{d}x$; and `FPR` is given by $\mathcal{C}_g^{\mathtt{FPR}}(\theta) = \int_{\theta_g}^\infty f_g^0(x)\mathrm{d}x$.

Accordingly, the firm's optimal choice of decision thresholds can be determined from:

$$\max_{\theta_a, \theta_b} \sum_{g \in \{a,b\}} n_g \int (\alpha_g u_+ f_g^1(x) - (1 - \alpha_g) u_- f_g^0(x)) \pi_g(x) \mathrm{d}x \quad \text{s.t.} \qquad \mathcal{C}_g^{\mathtt{f}}(\theta_a) = \mathcal{C}_g^{\mathtt{f}}(\theta_b)\,. \quad (1)$$

Let $\theta_g^{\mathtt{f}}$ denote the solution of (1) under fairness constraint $\mathtt{f} \in \{\mathtt{DP}, \mathtt{TPR}, \mathtt{FPR}\}$, and $\theta_g^{\mathtt{MU}}$ denote the `Maximum Utility (MU)` thresholds (i.e., unconstrained maximizer of the expected utility).

[1] We also consider `Equalized Odds (EO)` (Hardt et al., 2016) in our experiments, which requires that both TPR and FPR rates be equalized. Also, the choice of hard constraints is for theoretical convenience. In our experiments, we allow for soft constraints of the form $|\mathcal{C}^{\mathtt{f}}(\theta_a) - \mathcal{C}^{\mathtt{f}}(\theta_b)| < \epsilon$.

Data biases. In order to solve (1), the firm relies on historical information and training datasets to obtain estimates of the underlying population characteristics α_g, $f_g^y(x)$, and/or $\gamma_g(x)$. However, the estimated quantities may differ from the true population characteristics. We refer to the inaccuracies in these estimates as data bias, and focus on *qualification assessment biases*, reflected in the form of errors in the qualification profile estimates $\hat{\gamma}_g(x)$. This captures past errors or biases in decision making, including labeling biases, which lead the firm to have an incorrect estimate of an individual's likelihood of qualification following observing its feature and sensitive attribute. We note that such biases also affect the estimates $\hat{\alpha}_g$ and $\hat{f}_g^y(x)$. We investigate the impacts of such biases on the firm's expected payoff and its ability to meet the desired fairness criteria $\texttt{f}$, for different $\texttt{f}$.

3 Analytical Results

The unconstrained thresholds. We begin by characterizing the unconstrained thresholds $\theta_g^{\texttt{MU}}$ that maximize the firm's expected utility, and investigate the impacts of data biases on these thresholds.

Lemma 1 (Optimal MU thresholds)**.** *The thresholds* $\{\theta_a^{MU}, \theta_b^{MU}\}$ *satisfy* $\gamma_g(\theta_g^{MU}) = \frac{u_-}{u_+ + u_-}$.

Lemma 2 (Impact of data biases on MU thresholds and firm's utility)**.** *Let* θ_g^{MU} *and* $\hat{\theta}_g^{MU}$ *denote the optimal MU decision thresholds for group* g*, obtained given unbiased data and data with biases on group* b*, respectively. If* $\hat{\gamma}_b(\theta_b^{MU}) < \gamma_b(\theta_b^{MU})$*, the decision threshold on group* b *increases, i.e.,* $\hat{\theta}_b^{MU} > \theta_b^{MU}$*. The reverse holds if the inequalities above are reversed. In all these cases, the decisions on group* a *are unaffected, i.e.,* $\hat{\theta}_a^{MU} = \theta_a^{MU}$*, and firm's utility drops, i.e.,* $U(\theta_a^{MU}, \hat{\theta}_b^{MU}) < U(\theta_a^{MU}, \theta_b^{MU})$.

As intuitively expected, biases against the disadvantaged group lead to an increase in their disadvantage; the reverse is true if the group is perceived more favorably. We also highlight that the decisions on group a remain unaffected by any biases in group b's data. We next analyze how the coupling introduced between the group's decisions due to fairness criteria breaks this independence.

Fairness-constrained thresholds. We begin by characterizing these thresholds.

Lemma 3 (Optimal fair thresholds)**.** *The thresholds* $\{\theta_a^{\texttt{f}}, \theta_b^{\texttt{f}}\}$ *solving* (1) *under fairness constraint* $\mathcal{C}_a^{\texttt{f}}(\theta_a) = \mathcal{C}_b^{\texttt{f}}(\theta_b)$*,* $\texttt{f} \in \{DP, TPR, FPR\}$*, satisfy* $\sum_{g\in\{a,b\}} n_g \frac{\alpha_g u_+ f_g^1(\theta_g^{\texttt{f}}) - (1-\alpha_g) u_- f_g^0(\theta_g^{\texttt{f}})}{\partial C_g^{\texttt{f}}(\theta_g^{\texttt{f}})/\partial\theta} = 0$.

This characterization is similar to that obtained in (Zhang et al., 2020). It can be further simplified for fairness constraints $\texttt{f} \in \{\texttt{DP}, \texttt{TPR}, \texttt{FPR}\}$ (as show in Table 2 in the appendix).

Impacts of qualification assessment biases. We now analyze the robustness of different fairness criteria against biased training data, by considering biases that result in $\hat{\gamma}_b(x) = \beta\gamma_b(x), \forall x$, where $\beta \in (0, 1]$ is the underestimation rate. This can be viewed as label biases due to a prior decision maker/policy that had a probability β of correctly identifying qualified agents from group b.

Proposition 1. *Assume* $\hat{\gamma}_b(x) = \beta\gamma_b(x), \forall x$*, where* $\beta \in (0, 1]$*. Let* $\theta_g^{\texttt{f}}$ *and* $\hat{\theta}_g^{\texttt{f}}(\beta)$ *denote the optimal decision thresholds satisfying fairness constraint* $\texttt{f} \in \{DP, TPR, FPR\}$*, obtained from unbiased data and data with biases on group* b *given* β*, respectively. Then, (i)* $\hat{\theta}_g^{\texttt{f}}(\beta) \geq \theta_g^{\texttt{f}}$ *for both groups and for any of the three constraints, for all* β*. Further,* $\hat{\theta}_g^{\texttt{f}}(\beta)$ *is decreasing in* β*. (ii) The* DP *and* TPR *fairness criteria continue to be met, while* FPR *is violated, at their corresponding* $\{\hat{\theta}_a^{\texttt{f}}(\beta), \hat{\theta}_b^{\texttt{f}}(\beta)\}$*. (iii) The firm's utility is non-increasing under* $\texttt{f} \in \{DP, TPR\}$*, i.e.,* $U(\hat{\theta}_a^{\texttt{f}}(\beta), \hat{\theta}_b^{\texttt{f}}(\beta)) \leq U(\theta_a^{\texttt{f}}, \theta_b^{\texttt{f}})$*. (iv) The firm's utility may increase or decrease under* FPR*.*

We begin by noting two main differences of these findings with Lemma 2 in the unconstrained setting: (1) the biases in group b's data now lead to under-selection of *both* groups compared to the unbiased case. That is, as expected, the introduction of fairness constraints couples the groups in the impact of data biases as well. (2) Perhaps more interestingly, there exist scenarios in which the adoption of a fairness constraint *benefits* a firm facing biased qualification assessments. Note however that the fairness requirement is no longer satisfied in such scenarios.

In addition to these observations, Proposition 1 shows that the DP and TPR fairness criteria are *robust* to underestimation of qualification profiles, in that the obtained thresholds continue to satisfy

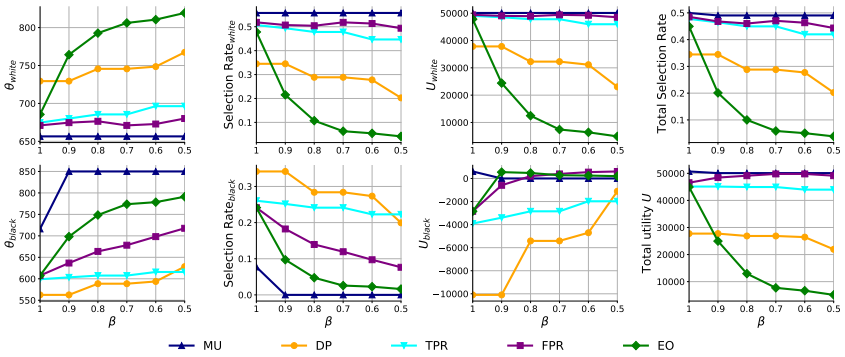

Figure 2: Thresholds, selection rates, and utility from numerical experiments on FICO data.

the desired notion of fairness.[2] That said, the changes in these thresholds lead to loss of utility for the firm. To better assess these impacts, we characterize the *sensitivity* of DP and TPR thresholds to qualification assessment errors (a detailed statement is given in Proposition 2 in the appendix). The following corollary states that DP is more sensitive to qualification assessment biases than TPR.

Corollary 1. *There, there exists a $\bar{\alpha}_b$ such that for all $\alpha_b \leq \bar{\alpha}_b$, we have* $|\frac{\partial \hat{\theta}_b^{TPR}(1)}{\partial \beta}| < |\frac{\partial \hat{\theta}_b^{DP}(1)}{\partial \beta}|$.

4 Numerical Experiments

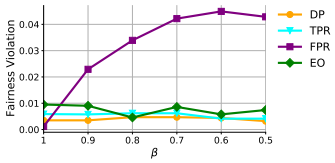

Figure 1: Fairness violation (FICO).

FICO credit score dataset. We first conduct experiments on the FICO dataset preprocessed by Hardt et al. (2016). The detials of our experiment settings are given in the appendix. To induce data biases, we take the repay probabilities of the black group in this dataset and and drop them by $\beta \in (0.5, 1)$ to model the underestimation of the qualification profiles. The firm's decision rules are found on this biased data and applied to the unbiased data. Figure 1 shows the fairness violation under each constraint with respect to the level of bias we induce in the data. We observe that, consistent with Proposition 1, DP and TPR are robust to underestimation of qualification profiles in terms of achieving their notions of fairness.

Figure 2 shows the changes in thresholds, selections rates, and the firm's utility (both overall and on each group). The thresholds for both groups increase as the bias level gets higher, which is in line with Proposition 1. In addition, the thresholds' increase under TPR is less drastic than DP and EO (consistent with Corollary 1). Lastly, we note the impact on the firm's utility. First note that due to the fact that the white group is the majority in this dataset, higher increase in θ_{white} leads to greater loss in total utility. We also note that when there is no bias, $\theta_a^{\text{DP}} > \theta_a^{\text{TPR}}$, and therefore the total utility under DP starts off much lower than that under TPR. It also drops faster, driven by the sharper increase in θ_a^{DP}. Moreover, the total utility may increase under FPR (consistent with Proposition 1).

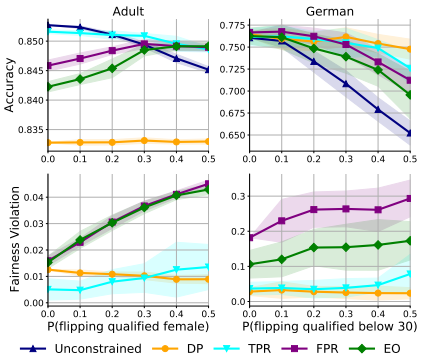

Figure 3: Accuracy and fairness violation on Adult and German datasets.

Adult dataset and German credit dataset. Figure 3 further evaluates the impacts of label flipping biases on the robustness of fairness constraints and on the classifier's accuracy on the UCI Adult and German credit datasets (Dua & Graff, 2017). We observe that similar conclusions hold in these datasets as well, with DP and TPR displaying robustness against label biases.

Acknowledgments. The authors are grateful for support from Cisco Research, and the NSF program on Fairness in AI in collaboration with Amazon under Award No. IIS-2040800. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, Amazon, or Cisco.

[2]Our experiments in Section 4 further support this observation by showing that these fairness criteria are more robust to label flipping biases in different real-world datasets.

REFERENCES

Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Steven Z Wu. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems*, pp. 8974–8984, 2019.

Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pp. 160–171. PMLR, 2018.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712, 2020.

Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 277–287. PMLR, 2020.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pp. 3150–3158. PMLR, 2018.

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

Seth Neel and Aaron Roth. Mitigating bias in adaptive data gathering via differential privacy. In *International Conference on Machine Learning*, pp. 3720–3729. PMLR, 2018.

Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics*, pp. 1261–1269, 2018.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *International Conference on Machine Learning*, pp. 11035–11046. PMLR, 2021.

Yifan Yang, Yang Liu, and Parinaz Naghizadeh. Adaptive data debiasing through bounded exploration and fairness. *arXiv preprint arXiv:2110.13054*, 2021.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

APPENDIX

A SUMMARY OF NOTATION

Notation	Description
g	demographic groups, $g \in \{a, b\}$
n_g	fraction of population from group g
x	observable feature, $x \in \mathbb{R}$
y	true qualification state, $y \in \{0, 1\}$
α_g	qualification rate of group g
$f_g^y(x)$	feature distribution of agents with label y from group g
$\gamma_g(x)$	qualification profile of group g; probability that agent with feature x from group g is qualified
d	firm's accept/reject decision, $d \in \{0, 1\}$
θ_g	firm's threshold policy on group g
u_+/u_-	firm's benefit/loss from accepting qualified/unqualified agents
$U(\theta_a, \theta_b)$	firm's expected payoff given policies $\{\theta_a, \theta_b\}$
$C^{\mathtt{f}}(\theta_g)$	fairness measure on group g
$\theta_g^{\mathtt{MU}}$	firm's optimal threshold on group g for maximum utility (fairness unconstrained)
$\theta_g^{\mathtt{f}}$	firm's optimal threshold on group g under fairness constraint $\mathtt{f}$

Table 1: Summary of notation.

B PROOFS FOR SECTION 3

We will use the following relation between the problem primitives in the proofs:

$$\gamma_g(x) = \frac{f_g^1(x)\alpha_g}{f_g^1(x)\alpha_g + f_g^0(x)(1-\alpha_g)} = \frac{1}{1 + \frac{f_g^0(x)}{f_g^1(x)}(\frac{1}{\alpha_g} - 1)} \,. \tag{2}$$

We note that under Assumption 1, $\gamma_g(x)$ is increasing in x.

B.1 PROOF OF LEMMA 1

Proof. In the absence of a fairness constraint, the firm's utility on each group, $U_g(\theta_g) = \int_{\theta_g}^{\infty}(\alpha_g u_+ f_g^1(x) - (1-\alpha_g)u_- f_g^0(x))\mathrm{d}x$ can be maximized independently. We have

$$\frac{\partial U_g(\theta_g)}{\partial \theta_g} = -(\alpha_g u_+ f_g^1(\theta_g) - (1-\alpha_g)u_- f_g^0(\theta_g)) \,.$$

By Assumption 1 and the above, $\alpha_g u_+ f_g^1(\theta_g^{\mathtt{MU}}) = (1-\alpha_g)u_- f_g^0(\theta_g^{\mathtt{MU}})$ is the maximizer of $U_g(\theta_g)$, establishing that $\frac{f_g^1(\theta_g^{\mathtt{MU}})}{f_g^0(\theta_g^{\mathtt{MU}})} = \frac{(1-\alpha_g)u_-}{\alpha_g u+}$. The expression in terms of the qualification profiles follows from (2). □

B.2 PROOF OF LEMMA 2

Proof. The increase in the decision thresholds on group b follows from the characterizations in Lemma 1, and noting that the functions $\frac{f_b^1(x)}{f_b^0(x)}$ and $\gamma_b(x)$ are both increasing. The thresholds on group a remain unaffected since the optimization of $U_a(\theta_a)$ and $U_b(\theta_b)$ can be decoupled. The reduction in the firm's utility follows from the proof of Lemma 1): $\{\theta_a^{\mathtt{MU}}, \theta_b^{\mathtt{MU}}\}$ is the only maximizer of the firm's utility, and so the utility drops at the perturbed thresholds. □

B.3 PROOF OF LEMMA 3

Proof. Given threshold policies, the firm's problem in (1) can be further simplified to

$$\max_{\theta_a,\theta_b} \sum_{g\in\{a,b\}} n_g \left(\alpha_g u_+(1-F_g^1(\theta_g)) - (1-\alpha_g)u_-(1-F_g^0(\theta_g))\right)$$
$$\text{s.t.} \quad \mathcal{C}_a^{\mathtt{f}}(\theta_a) = \mathcal{C}_b^{\mathtt{f}}(\theta_b)\ .$$

We first use the constraint to express θ_a as a function of θ_b, that is, find a function $\phi^{\mathtt{f}}$ such that $\theta_a = \phi^{\mathtt{f}}(\theta_b)$. To do so, note that for all three constraints $\mathtt{f} \in \{\mathtt{DP}, \mathtt{TPR}, \mathtt{FPR}\}$, the function $C_g^{\mathtt{f}}$ is invertible. Therefore, $\theta_a = (C_a^{\mathtt{f}})^{-1}(C_b^{\mathtt{f}}(\theta_b))$. Further,

$$\frac{\partial \phi^{\mathtt{f}}(\theta_b)}{\partial \theta_b} = \frac{\partial (C_a^{\mathtt{f}})^{-1}(C_b^{\mathtt{f}}(\theta_b))}{\partial \theta_b} = \frac{1}{\frac{\partial C_a^{\mathtt{f}}\left((C_a^{\mathtt{f}})^{-1}(C_b^{\mathtt{f}}(\theta_b))\right)}{\partial \theta}} \frac{\partial C_b^{\mathtt{f}}(\theta_b)}{\partial \theta} = \frac{\frac{\partial C_b^{\mathtt{f}}(\theta_b)}{\partial \theta}}{\frac{\partial C_a^{\mathtt{f}}(\theta_a)}{\partial \theta}} \tag{3}$$

It is also easy to show that $\frac{\partial C_g^{\mathtt{DP}}(\theta)}{\partial \theta} = \alpha_g f_g^1(\theta) + (1-\alpha_g) f_g^0(\theta)$, $\frac{\partial C_g^{\mathtt{TPR}}(\theta)}{\partial \theta} = f_g^1(\theta)$, and $\frac{\partial C_g^{\mathtt{FPR}}(\theta)}{\partial \theta} = f_g^0(\theta)$; therefore, $\phi^{\mathtt{f}}(\cdot)$ is an increasing function for all three constraints.

With the conversion $\theta_a = \phi^{\mathtt{f}}(\theta_b)$, the firm's optimizaiton problem reduces to

$$\begin{aligned}\theta_b^* = \arg\max_{\theta_b}\ & n_a\Big(\alpha_a u_+(1-F_a^1(\phi^{\mathtt{f}}(\theta_b))) \\ & - (1-\alpha_a)u_-(1-F_a^0(\phi^{\mathtt{f}}(\theta_b)))\Big) \\ & + n_b\Big(\alpha_b u_+(1-F_b^1(\theta_b)) - (1-\alpha_b)u_-(1-F_b^0(\theta_b))\Big)\end{aligned}$$

The first derivative of the objective function above with respect to θ_b is given by

$$\begin{aligned}\frac{\partial U(\phi^{\mathtt{f}}(\theta_b),\theta_b)}{\partial \theta_b} = & - n_a \frac{\frac{\partial C_b^{\mathtt{f}}(\theta_b)}{\partial \theta}}{\frac{\partial C_a^{\mathtt{f}}(\theta_a)}{\partial \theta}}\Big(\alpha_a u_+ f_a^1(\theta_a) - (1-\alpha_a)u_- f_a^0(\theta_a)\Big) \\ & - n_b\Big(\alpha_b u_+ f_b^1(\theta_b) - (1-\alpha_b)u_- f_b^0(\theta_b)\Big)\end{aligned}$$

Consider the thresholds $\{\theta_a^*, \theta_b^*\}$ at which this first derivative is zero. Given that $\phi^{\mathtt{f}}$ is increasing, alternative feasible profiles have either both thresholds smaller or larger than those in $\{\theta_a^*, \theta_b^*\}$. Together with Assumption 1, this first derivative is positive/negative when the thresholds decrease/increase. We therefore conclude that $\{\theta_a^*, \theta_b^*\}$ maximizes the firm's utility.

That is, the optimal fairness-constrained thresholds satisfy

$$n_a \frac{\frac{\partial C_b^{\mathtt{f}}(\theta_b^{\mathtt{f}})}{\partial \theta}}{\frac{\partial C_a^{\mathtt{f}}(\theta_a^{\mathtt{f}})}{\partial \theta}}\Big(\alpha_a u_+ f_a^1(\theta_a^{\mathtt{f}}) - (1-\alpha_a)u_- f_a^0(\theta_a^{\mathtt{f}})\Big) + n_b\Big(\alpha_b u_+ f_b^1(\theta_b^{\mathtt{f}}) - (1-\alpha_b)u_- f_b^0(\theta_b^{\mathtt{f}})\Big) = 0$$

completing the proof. □

B.4 OPTIMAL FAIRNESS-CONSTRAINED THRESHOLDS UNDER DIFFERENT CHOICES OF F

Using Lemma 3, we can further characterize the thresholds for fairness criteria $\mathtt{f} \in \{\mathtt{DP}, \mathtt{TPR}, \mathtt{FPR}\}$, as shown in Table 2.

B.5 DERIVATIONS OF THRESHOLDS IN TABLE 2

It is straightforward to verify that $\frac{\partial C_g^{\mathtt{DP}}(\theta)}{\partial \theta} = \alpha_g f_g^1(\theta) + (1-\alpha_g) f_g^0(\theta)$, $\frac{\partial C_g^{\mathtt{TPR}}(\theta)}{\partial \theta} = f_g^1(\theta)$, and $\frac{\partial C_g^{\mathtt{FPR}}(\theta)}{\partial \theta} = f_g^0(\theta)$. We next derive the characterizations in Table 2 from this together with the relation in (2). Specifically,[3]

[3]We will use $\pm$ as a shorthand for a term being added and subtracted.

f	Optimal thresholds in terms of $\gamma_g(x)$ and α_g
DP	$n_a\gamma_a(\theta_a^{\texttt{DP}}) + n_b\gamma_b(\theta_b^{\texttt{DP}}) = \frac{u_-}{u_+ + u_-}$
TPR	$\frac{n_a\alpha_a}{\gamma_a(\theta_a^{\texttt{TPR}})} + \frac{n_b\alpha_b}{\gamma_b(\theta_b^{\texttt{TPR}})} = \frac{1}{\frac{u_-}{u_+ + u_-}}(n_a\alpha_a + n_b\alpha_b)$
FPR	$\frac{n_a(1-\alpha_a)}{1-\gamma_a(\theta_a^{\texttt{FPR}})} + \frac{n_b(1-\alpha_b)}{1-\gamma_b(\theta_b^{\texttt{FPR}})} = \frac{1}{1-\frac{u_-}{u_+ + u_-}}(n_a(1-\alpha_a) + n_b(1-\alpha_b))$

Table 2: Optimal fairness-constrained thresholds under different choices of $\texttt{f} \in \{\texttt{DP}, \texttt{TPR}, \texttt{FPR}\}$.

- For DP:

$$
\begin{aligned}
&\sum_{g\in\{a,b\}} n_g \frac{\alpha_g u_+ f_g^1(\theta_g^{\texttt{DP}}) - (1-\alpha_g)u_- f_g^0(\theta_g^{\texttt{DP}})}{\alpha_g f_g^1(\theta_g^{\texttt{DP}}) + (1-\alpha_g) f_g^0(\theta_g^{\texttt{DP}})} = 0 \\
\Leftrightarrow\ &\sum_{g\in\{a,b\}} n_g \frac{\alpha_g u_+ f_g^1(\theta_g^{\texttt{DP}}) - (1-\alpha_g)u_- f_g^0(\theta_g^{\texttt{DP}}) \pm \alpha_g u_- f_g^1(\theta_g^{\texttt{DP}})}{\alpha_g f_g^1(\theta_g^{\texttt{DP}}) + (1-\alpha_g) f_g^0(\theta_g^{\texttt{DP}})} = 0 \\
\Leftrightarrow\ &\sum_{g\in\{a,b\}} n_g \left(\frac{\alpha_g f_g^1(\theta_g^{\texttt{DP}})}{\alpha_g f_g^1(\theta_g^{\texttt{DP}}) + (1-\alpha_g) f_g^0(\theta_g^{\texttt{DP}})}(u_+ + u_-) - u_- \right) = 0 \\
\Leftrightarrow\ &\sum_{g\in\{a,b\}} n_g \gamma_g(\theta_g^{\texttt{DP}}) = \sum_{g\in\{a,b\}} n_g \frac{u_-}{u_+ + u_-} \\
\Leftrightarrow\ &n_a\gamma_a(\theta_a^{\texttt{DP}}) + n_b\gamma_b(\theta_b^{\texttt{DP}}) = \frac{u_-}{u_+ + u_-}
\end{aligned}
$$

- For TPR:

$$
\begin{aligned}
&\sum_{g\in\{a,b\}} n_g \left(\alpha_g u_+ - (1-\alpha_g)u_- \frac{f_g^0(\theta_g^{\texttt{TPR}})}{f_g^1(\theta_g^{\texttt{TPR}})} \right) = 0 \\
\Leftrightarrow\ &\sum_{g\in\{a,b\}} n_g \frac{\alpha_g u_+ f_g^1(\theta_g^{\texttt{TPR}}) - (1-\alpha_g)u_- f_g^0(\theta_g^{\texttt{TPR}})}{f_g^1(\theta_g^{\texttt{TPR}})} = 0 \\
\Leftrightarrow\ &\sum_{g\in\{a,b\}} n_g \frac{\alpha_g u_+ f_g^1(\theta_g^{\texttt{TPR}}) - (1-\alpha_g)u_- f_g^0(\theta_g^{\texttt{TPR}}) \pm \alpha_g u_- f_g^1(\theta_g^{\texttt{TPR}})}{f_g^1(\theta_g^{\texttt{TPR}})} = 0 \\
\Leftrightarrow\ &\sum_{g\in\{a,b\}} n_g \left(\alpha_g(u_+ + u_-) - u_- \frac{\alpha_g f_g^1(\theta_g^{\texttt{TPR}}) + (1-\alpha_g) f_g^0(\theta_g^{\texttt{TPR}})}{f_g^1(\theta_g^{\texttt{TPR}})} \right) = 0 \\
\Leftrightarrow\ &\sum_{g\in\{a,b\}} n_g \left(\alpha_g \frac{u_+ + u_-}{u_-} - \frac{\alpha_g}{\gamma_g(\theta_g^{\texttt{TPR}})} \right) = 0 \\
\Leftrightarrow\ &\frac{n_a\alpha_a}{\gamma_a(\theta_a^{\texttt{TPR}})} + \frac{n_b\alpha_b}{\gamma_b(\theta_b^{\texttt{TPR}})} = \frac{1}{\frac{u_-}{u_+ + u_-}}(n_a\alpha_a + n_b\alpha_b)
\end{aligned}
$$

- For FPR:

$$\sum_{g\in\{a,b\}} n_g \left(\alpha_g u_+ \frac{f_g^1(\theta_g^{\texttt{FPR}})}{f_g^0(\theta_g^{\texttt{FPR}})} - (1-\alpha_g)u_-\right) = 0$$

$$\Leftrightarrow \sum_{g\in\{a,b\}} n_g \frac{\alpha_g u_+ f_g^1(\theta_g^{\texttt{FPR}}) - (1-\alpha_g)u_- f_g^0(\theta_g^{\texttt{FPR}})}{f_g^0(\theta_g^{\texttt{FPR}})} = 0$$

$$\Leftrightarrow \sum_{g\in\{a,b\}} n_g \frac{\alpha_g u_+ f_g^1(\theta_g^{\texttt{FPR}}) - (1-\alpha_g)u_- f_g^0(\theta_g^{\texttt{FPR}}) \pm (1-\alpha_g)u_+ f_g^0(\theta_g^{\texttt{FPR}})}{f_g^0(\theta_g^{\texttt{FPR}})} = 0$$

$$\Leftrightarrow \sum_{g\in\{a,b\}} n_g \left(u_+ \frac{\alpha_g f_g^1(\theta_g^{\texttt{FPR}}) + (1-\alpha_g) f_g^0(\theta_g^{\texttt{FPR}})}{f_g^0(\theta_g^{\texttt{FPR}})} - (1-\alpha_g)(u_+ + u_-)\right) = 0$$

$$\Leftrightarrow \sum_{g\in\{a,b\}} n_g \left(\frac{(1-\alpha_g)}{1-\gamma_g(\theta_g^{\texttt{FPR}})} - (1-\alpha_g)\frac{u_+ + u_-}{u_+}\right) = 0$$

$$\Leftrightarrow \frac{n_a(1-\alpha_a)}{1-\gamma_a(\theta_a^{\texttt{FPR}})} + \frac{n_b(1-\alpha_b)}{1-\gamma_b(\theta_b^{\texttt{FPR}})} = \frac{1}{1-\frac{u_-}{u_+ + u_-}}(n_a(1-\alpha_a) + n_b(1-\alpha_b))$$

B.6 Proof of Proposition 1

Proof. We begin with some preliminaries, determining the impacts of $\hat{\gamma}_b(x) = \beta\gamma_b(x)$ on other estimates of the underlying population characteristics. To do so, let $f_g(x) := \mathbb{P}(X = x|G = g) = \alpha_g f_g^1(x) + (1-\alpha_g) f_g^0(x)$ be the feature distribution across all agents (qualified and unqualified) from group g. We note that if there is a bias in qualification assessments $\hat{\gamma}_g(x) = \beta\gamma_g(x)$(which affects the labels y), these overall feature distribution estimates $\hat{f}_g(x) = \mathbb{P}(X = x|G = g)$ will not be affected; that is $\hat{f}_g(x) = f_g(x)$. However, $\hat{\alpha}_g$ and $\hat{f}_g^y(x)$ can still be impacted.

We first identify the impacts of the change in $\gamma_g(x)$ on α_g. By definition:

$$\alpha_g = \int_x \mathbb{P}(Y = 1|X = x, G = g)\mathbb{P}(X = x|G = g)\mathrm{d}x \,.$$

Therefore, if $\hat{\gamma}_b(x) = \beta\gamma_b(x), \forall x$, we have $\hat{\alpha}_b = \beta\alpha_b$.

Now, by definition and the Bayes' rule

$$f_g^y(x) = \mathbb{P}(X = x|Y = y, G = g) = \frac{\mathbb{P}(Y=y|X=x,G=g)\mathbb{P}(X=x|G=g)}{\mathbb{P}(Y=y|G=g)} \,.$$

Therefore,

$$f_g^1(x) = \frac{\gamma_g(x) f_g(x)}{\alpha_g} \,.$$
$$f_g^0(x) = \frac{(1-\gamma_g(x)) f_g(x)}{(1-\alpha_g)} \,.$$

Noting that $\hat{\alpha} = \beta_g\alpha_g$, combined with the above relations, we conclude that $\hat{f}_g^1(x) = f_g^1(x)$, while $\hat{f}_g^0(x) = \frac{1-\alpha_g}{1-\beta\alpha_g}\frac{1-\beta\gamma_g(x)}{1-\gamma_g(x)} f_g^0(x)$. Intuitively, this is expected: $\hat{\gamma}_g(x) = \beta\gamma_g(x)$ can be viewed as flipping label 1 to label 0 in the training data with probability β. This leaves the feature distribution of qualified agents unchanged (as the flipping probability is independent of the feature x), whereas it adds (incorrect) data to the feature distribution of unqualified agents, hence biasing $f_g^0(x)$.

We now proceed with the proof of the proposition.

Part (i):

• For DP: The firm picks the thresholds such that $h^{\texttt{DP}}(\hat{\theta}_a^{\texttt{DP}}, \hat{\theta}_b^{\texttt{DP}}, \beta) = 0$, where

$$h^{\texttt{DP}}(\theta_a, \theta_b, \beta) := n_a\gamma_a(\theta_a) + n_b\beta\gamma_b(\theta_b) - \frac{u_-}{u_+ + u_-} \,.$$

Note that as $\gamma_g(x)$ are increasing functions under Assumption 1, $h^{\texttt{DP}}$ is increasing in both thresholds. It is also increasing in β.

For $\beta \in (0,1)$, $h^{\texttt{DP}}(\theta_a^{\texttt{DP}}, \theta_b^{\texttt{DP}}, \beta) < 0$. Therefore, to attain $h^{\texttt{DP}}(\hat{\theta}_a^{\texttt{DP}}, \hat{\theta}_b^{\texttt{DP}}, \beta) = 0$, at least one of the thresholds should increase compared to the unbiased thresholds $\{\theta_a^{\texttt{DP}}, \theta_b^{\texttt{DP}}\}$. In addition, the DP-constrained thresholds when data is biased are selected such that $\int_{\hat{\theta}_a^{\texttt{DP}}}^{\infty} f_a(x)\mathrm{d}x = \int_{\hat{\theta}_b^{\texttt{DP}}}^{\infty} \hat{f}_b(x)\mathrm{d}x$ (i.e., based on the biased training data); since $\hat{f}_b(x) = f_b(x)$, we conclude that the changes in the thresholds are aligned (i.e, either both decrease or both increase compared to the unbiased case). We conclude that $\hat{\theta}_g^{\texttt{DP}}(\beta) \geq \theta_g^{\texttt{DP}}$ for both groups.

• For TPR: from Table 2, The firm picks the thresholds such that $h^{\texttt{TPR}}(\hat{\theta}_a^{\texttt{TPR}}, \hat{\theta}_b^{\texttt{TPR}}, \beta) = 0$, where

$$h^{\texttt{TPR}}(\theta_a, \theta_b, \beta) := \frac{n_a\alpha_a}{\gamma_a(\theta_a^{\texttt{TPR}})} + \frac{n_b\alpha_b}{\gamma_b(\theta_b^{\texttt{TPR}})} - \frac{1}{\frac{u_-}{u_+ + u_-}}(n_a\alpha_a + n_b\beta\alpha_b)\ .$$

Note that as $\gamma_g(x)$ are increasing functions under Assumption 1, $h^{\texttt{TPR}}$ is decreasing in both thresholds. It is also decreasing in β.

For $\beta \in (0,1)$, $h^{\texttt{TPR}}(\theta_a^{\texttt{TPR}}, \theta_b^{\texttt{TPR}}, \beta) > 0$. Therefore, to attain $h^{\texttt{TPR}}(\hat{\theta}_a^{\texttt{TPR}}, \hat{\theta}_b^{\texttt{TPR}}, \beta) = 0$, at least one of the thresholds should increase compared to the unbiased thresholds $\{\theta_a^{\texttt{TPR}}, \theta_b^{\texttt{TPR}}\}$. In addition, as $\hat{f}_g^1(x) = f_g^1(x)$, and by the definition of TPR, the changes in the thresholds are aligned (i.e, either both decrease or both increase compared to the unbiased case). Therefore, $\hat{\theta}_g^{\texttt{TPR}}(\beta) \geq \theta_g^{\texttt{TPR}}$ for both groups.

• For FPR: based on Table 2, The firm picks the thresholds such that $h^{\texttt{FPR}}(\hat{\theta}_a^{\texttt{FPR}}, \hat{\theta}_b^{\texttt{FPR}}, \beta) = 0$, where

$$\begin{aligned} h^{\texttt{FPR}}(\theta_a, \theta_b, \beta) :=& \tfrac{n_a(1-\alpha_a)}{1-\gamma_a(\theta_a)} + \tfrac{n_b(1-\beta\alpha_b)}{1-\beta\gamma_b(\theta_b)} \\ & - \frac{1}{1 - \frac{u_-}{u_+ + u_-}}(n_a(1-\alpha_a) + n_b(1-\beta\alpha_b)) \end{aligned}$$

Note that as $\gamma_g(x)$ are increasing functions under Assumption 1, $h^{\texttt{TPR}}$ is increasing in both thresholds. To identify its trend in β, the derivative of $h^{\texttt{FPR}}$ with respect to β is given by

$$\begin{aligned} \frac{\partial h^{\texttt{FPR}}}{\partial \beta} =& n_b \frac{-\alpha_b(1-\beta\gamma_b(\theta_b)) + \gamma_b(\theta_b)(1-\beta\alpha_b)}{(1-\beta\gamma_b(\theta_b))^2} + \frac{n_b\alpha_b}{1 - \frac{u_-}{u_+ + u_-}} \\ =& \alpha_b n_b \left(\frac{1}{1 - \frac{u_-}{u_+ + u_-}} - \frac{1}{1-\beta\gamma_b(\theta_b)} \right) + \frac{n_b\gamma_b(\theta_b)(1-\beta\alpha_b)}{(1-\beta\gamma_b(\theta_b))^2} \end{aligned}$$

As $h^{\texttt{FPR}}(\hat{\theta}_a^{\texttt{FPR}}, \hat{\theta}_b^{\texttt{FPR}}, \beta) = 0$, and $\gamma_b(x) \leq \gamma_a(x)$, we can conclude that $\frac{1}{1-\beta\gamma_b(\hat{\theta}_b^{\texttt{FPR}})} \leq \frac{1}{1-\frac{u_-}{u_+ + u_-}} \leq \frac{1}{1-\gamma_a(\hat{\theta}_a^{\texttt{FPR}})}$. This means that $\frac{\partial h^{\texttt{FPR}}}{\partial \beta}$ is increasing at the optimal biased thresholds for each $\beta \in (0,1)$. Therefore, $h^{\texttt{FPR}}(\theta_a^{\texttt{FPR}}, \theta_b^{\texttt{FPR}}, \beta) \leq 0$ as β decreases from 1. That means that in order to attain $h^{\texttt{FPR}}(\hat{\theta}_a^{\texttt{FPR}}, \hat{\theta}_b^{\texttt{FPR}}, \beta) = 0$, at least one of the thresholds should increase compared to the unbiased thresholds $\{\theta_a^{\texttt{FPR}}, \theta_b^{\texttt{FPR}}\}$. In addition, from (3) in the proof of Lemma 3, we know that given that $\beta = 1$, if $\theta_a^{\texttt{FPR}}$ drops, so should $\theta_b^{\texttt{FPR}}$ for the FPR constraint to continue to hold So it must be that both thresholds increase. We conclude that both thresholds should increase, that is $\hat{\theta}_g^{\texttt{FPR}}(\beta) \geq \theta_g^{\texttt{FPR}}$ for both groups when β drops from 1.

Lastly, for all three constraints, applying the same argument to levels of qualification assessment biases $\beta^1 > \beta^2$, we conclude that $\hat{\theta}_g^{\texttt{f}}(\beta^2) > \hat{\theta}_g^{\texttt{f}}(\beta^1)$. That is, $\hat{\theta}_g^{\texttt{f}}(\beta)$ is decreasing in β.

Part (ii): The DP-constrained thresholds when data is biased are selected such that $\int_{\hat{\theta}_a^{\texttt{DP}}}^{\infty} f_a(x)\mathrm{d}x = \int_{\hat{\theta}_b^{\texttt{DP}}}^{\infty} \hat{f}_b(x)\mathrm{d}x$ (i.e., based on the biased training data); since $\hat{f}_b(x) = f_b(x)$, this constraint is also satisfied on the unbiased data. That is, DP continues to hold (on the unbiased training data, as intended) at $\{\hat{\theta}_a^{\texttt{DP}}, \hat{\theta}_b^{\texttt{DP}}\}$.

Next, as $\hat{f}_g^1(x) = f_g^1(x)$ and $\hat{f}_b^0(x) = \frac{1-\alpha_b}{1-\beta\alpha_b}\frac{1-\beta\gamma_b(x)}{1-\gamma_b(x)} f_b^0(x)$, the thresholds satisfying TPR on the biased data also satisfy TPR on the unbiased data, while the same is not true for FPR.

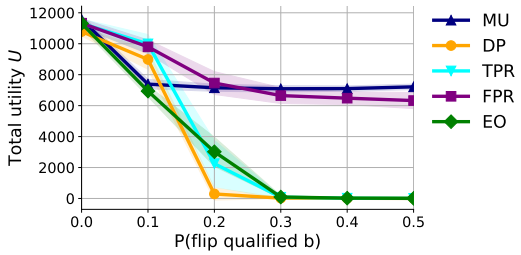

Figure 4: Total utility for case where the firm's expected utility decreases under FPR. The results are averaged over 20 runs.

Part (iii): We note that by the previous part, the new thresholds $\{\hat{\theta}_a^{\mathtt{f}}(\beta), \hat{\theta}_b^{\mathtt{f}}(\beta)\}$ continue to satisfy the DP and TPR constraints. That means that these constraints were also feasible choices in the unbiased scenario. Therefore, $U(\hat{\theta}_a^{\mathtt{f}}(\beta), \hat{\theta}_b^{\mathtt{f}}(\beta)) \leq U(\theta_a^{\mathtt{f}}, \theta_b^{\mathtt{f}})$ as $\{\theta_a^{\mathtt{f}}, \theta_b^{\mathtt{f}}\}$ are the expected utility maximizers among all feasible threshold choices in the unconstrained case.

Part (iv): We prove this part by construction. An example for utility drop under FPR is given in Figure 4. An example for utility increase under FPR is given Figure 2 in our numerical experiments in Section 4.

To provide additional intuition for this behavior, note that the firm's expected utility on group g can be written as

$$U_g(\theta) = \int_{\theta}^{\infty} [\gamma_g(x)(u_+ + u_-) - u_-] f_g(x) \mathrm{d}x \ .$$

Therefore, if $\gamma_g(x) \geq \frac{u_-}{u_+ + u_-}$, the utility is decreasing in the threshold, and vice versa.

From the proof of part (i), we know that $\frac{1}{1-\beta\gamma_b(\hat{\theta}_b^{\mathtt{FPR}})} \leq \frac{1}{1-\frac{u_-}{u_+ + u_-}} \leq \frac{1}{1-\gamma_a(\hat{\theta}_a^{\mathtt{FPR}})}$. Therefore, $\gamma_a(\hat{\theta}_a^{\mathtt{FPR}}) \geq \frac{u_-}{u_+ + u_-}$, and as the threshold on group a increases given the increase in $\hat{\theta}_a^{\mathtt{FPR}}$ with increasing noise, the utility that the firm derives from group a drops. On the other hand, given $\frac{1}{1-\beta\gamma_b(\hat{\theta}_b^{\mathtt{FPR}})} \leq \frac{1}{1-\frac{u_-}{u_+ + u_-}}$, we know that $\beta\gamma_b(\hat{\theta}_b^{\mathtt{FPR}}) \leq \frac{u_-}{u_+ + u_-}$. However, for sufficiently small β, it might be that the above holds, while $\gamma_b(\hat{\theta}_b^{\mathtt{FPR}}) \geq \frac{u_-}{u_+ + u_-}$. Therefore, the utility from group b may increase at large β (in a way that the firm's overall utility increases), and may then drop at sufficiently small β (so that the firm's overall utility decreases). □

B.7 Sensitivity of DP / TPR thresholds to qualification assessment biases

Proposition 2 (Sensitivity of DP / TPR thresholds to qualification assessment biases)**.** *Consider the same setting as Proposition 1. Then, the rate of change of group b's thresholds at $\beta = 1$ is given by*

$$\frac{\partial \hat{\theta}_b^{DP}(\beta)}{\partial \beta}|_{\beta=1} = -\frac{1}{\frac{n_a}{n_b}\frac{f_b(\theta_b^{DP})}{f_a(\theta_a^{DP})}\frac{\gamma_a'(\theta_a^{DP})}{\gamma_b(\theta_b^{DP})} + \frac{\gamma_b'(\theta_b^{DP})}{\gamma_b(\theta_b^{DP})}}$$

$$\frac{\partial \hat{\theta}_b^{TPR}(\beta)}{\partial \beta}|_{\beta=1} = -\frac{1 + \frac{u_+}{u_-}}{\frac{n_a}{n_b}\frac{\alpha_a}{\alpha_b}\frac{f_b^1(\theta_b^{TPR})}{f_a^1(\theta_a^{TPR})}\frac{\gamma_a'(\theta_a^{TPR})}{(\gamma_a(\theta_a^{TPR}))^2} + \frac{\gamma_b'(\theta_b^{TPR})}{(\gamma_b(\theta_b^{TPR}))^2}}$$

Further, the rate of change of group a's thresholds at $\beta = 1$ is given by

$$\frac{\partial \hat{\theta}_a^{DP}(\beta)}{\partial \beta}|_{\beta=1} = \frac{f_b(\theta_b^{DP})}{f_a(\theta_a^{DP})}\frac{\partial \hat{\theta}_b^{DP}(\beta)}{\partial \beta}|_{\beta=1}$$

$$\frac{\partial \hat{\theta}_a^{TPR}(\beta)}{\partial \beta}|_{\beta=1} = \frac{f_b^1(\theta_b^{TPR})}{f_a^1(\theta_a^{TPR})}\frac{\partial \hat{\theta}_b^{TPR}(\beta)}{\partial \beta}|_{\beta=1}$$

The proof follows from the characterizations in Table 2, as well as the proof of Lemma 3, particularly (3).

As consistent with Proposition 1, this proposition shows that the thresholds increase when qualification assessments become biased (i.e., β decreases from 1). We see that under the DP fairness constraint, the increase in group b's threshold is higher if $\frac{n_a}{n_b}$ is smaller (i.e., the representation of group b increases). For the TRP fairness constraint on the other hand, the increase in group b's threshold is impacted by additional problem parameters: it is higher if $\frac{n_a}{n_b}$ or $\frac{\alpha_a}{\alpha_b}$ is smaller (i.e., the representation or qualification rate of group b increases), or if $\frac{u_+}{u_-}$ is larger (i.e., the firm's benefit/loss increases/decreases). In addition, under the DP threshold, the drop in group a's threshold is smaller if the qualification gap between the two groups is larger.

C Details on the numerical experiments

In this section, we detail our experimental setup on the FICO dataset preprocessed by Hardt et al. (2016), and provide additional interpretations for our results.

The FICO credit scores ranging from 300 to 850 correspond to the one-dimensional feature x in our model, and race is the sensitive feature that defines g. The repay probability for each score and group matches our qualification profile $\gamma_g(x)$. $\frac{u_-}{u_+}$ is set to 10. We focus on the black and white groups in our discussion. Additionally, n_{black} and n_{white} are 0.12 and 0.88, and α_{black} and α_{white} are 0.34 and 0.76.

We take the repay probabilities of the black group and drop them to model the underestimation of the qualification profiles. Decision rules will be found on the biased data and applied to the unbiased data. [4]

We have analyzed the effects of bias with and without the presence of fairness constraints. We measure the fairness violation under each constraint based on its fairness definition with respect to the level of bias we impose on the data. We have also observed the change in the thresholds, utility, and selection rates of each group.

From Figure 1, DP and TPR are robust to underestimation of qualification profiles in terms of achieving their notions of fairness. Based on the definition of DP, the constraint tries to equalize selection rates of two groups regardless of the true qualification state of the agents. Thus, each pair of decision rules that can achieve DP on the original unbiased data will still be able to achieve DP on the biased data. Conversely, the decision rules found on the biased data satisfying DP remain fair when applied back to the unbiased data.

Recall the definition of true positive rate, which is the ratio of the number of accepted qualified agents to the total number of qualified agents. Underestimating the qualification profiles would be equivalent to dropping the number of qualified agents at each score. When calculating the true positive rate, we decrease the two quantities in the ratio by the same fraction. Consequently, similar to DP, the set of decision rules satisfying TPR remains the same after the bias is imposed.

For EO, since it combines both TPR and FPR, the set of decision rules achieving EO will be at most the intersection of those of TPR and FPR. Figure 1 shows that FPR has an increasing trend in fairness violation. It means that the set of possible decision rules of FPR changes when bias is imposed. As false positive rate computes the ratio of the number of accepted unqualified agents to the total number of unqualified agents, dropping the qualification profiles increases the two quantities in the ratio at different rates, causing the false positive rates to change at given thresholds. As a result, the set of possible decision rules of EO will be different in size with bias but can still possibly contain decision rules close to certain ones, though may not be optimal in terms of utility, in the original unbiased set.

In Figure 2, we display the thresholds change of each group. The thresholds for both groups increase as the bias level gets higher, which is in line with Proposition 1. In addition, the thresholds increase under TPR is less drastic than DP and EO. The comparison of sensitivity of DP and TPR is consistent with what Corollary 1 suggests. We argue that the set of possible thresholds for EO changes when bias is imposed. In other words, for every bias level, EO searches a different set, making the thresholds change more significant.

[4] Unless otherwise specified, we use soft constraints $|\mathcal{C}^{\mathrm{f}}(\theta_a) - \mathcal{C}^{\mathrm{f}}(\theta_b)| \leq 0.01$ when finding decision rules in our experiments.