



<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *Thirty-fourth Conference on Neural Information Processing Systems*.

Citation for the original published paper:

Tu, R., Zhang, X., Liu, Y., Kjellström, H., Liu, M. et al. (2020)

How Do Fair Decisions Fare in Long-term Qualification?

In:

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-283120>

How Do Fair Decisions Fare in Long-term Qualification?

Xueru Zhang*

University of Michigan, AnnArbor, USA
xueru@umich.edu

Ruibo Tu*

KTH Royal Institute of Technology
ruibo@kth.se

Yang Liu

Harvard University
YangL@seas.harvard.edu

Mingyan Liu

University of Michigan, AnnArbor, USA
mingyan@umich.edu

Hedvig Kjellström

KTH Royal Institute of Technology
email

Kun Zhang

Carnegie Mellon University
kunz1@cmu.edu

Cheng Zhang

Microsoft Research, Cambridge
Cheng.Zhang@microsoft.com

Abstract

Although many fairness criteria have been proposed for decision making, their long-term impact on the well-being of a population remains unclear. In this work, we study the dynamics of population qualification and algorithmic decisions under a partially observed decision making setting. By characterizing the equilibrium of such dynamics, we theoretically analyze the long-term impact of static fairness constraints on the equality and improvement of group well-being. Our results show that static fairness constraints can either promote the equality or exacerbate the disparity depending on the driven factor of qualification transitions and the effect of sensitive attributes on feature distributions. In turn, we consider possible effective interventions that improve group qualification or promote the equality of group qualification. Our theoretical results and experiments on static real-world datasets with simulated dynamics show the consistent findings with social science studies.

1 Introduction

Automated decision making systems that are trained with real-world data can have inherent bias and exhibit discrimination against disadvantaged groups. One common approach to alleviating the issue is to enforce fairness constraints upon the decision making process such that certain statistical measures (e.g., true positive rate, positive classification rate, etc.) across multiple groups are (approximately) equalized. However, their effectiveness has been studied mostly in a static framework, where only the immediate impact of the constraints is assessed but not their long-term consequences. Recent studies have shown that imposing static fairness criteria that intend to protect disadvantaged groups can actually lead to pernicious long-term effects [32, 46]. The long-term effects are mainly shaped by the interplay between algorithmic decisions and individuals' reactions [33], i.e., algorithmic decisions directly lead to changes in the underlying feature distribution, which then feeds back into the decision making process. Understanding how this type of coupled dynamics evolve is still a challenge [9].

Towards this end, we consider a dynamical system involving a discrete-time sequential decision making process and underlying population responses, at each time step, to the decisions being made; this is captured by changes of their features in the next time step. Our goal is to understand how (static) fairness criteria in decision making affect the evolution of group well-being and characterize equilibrium states of the system. We focus on *myopic policies* that maximize the immediate utility under static fairness constraints and then examine their impacts on different groups in the long run.

More specifically, we seek to study the dynamics of group qualification rates [29, 33, 36, 44] and evaluate the long-term impact of various static fairness constraints imposed on a binary decision made at each discrete time step. We wonder whether these static fairness can also mitigate the qualification disparity in the long-run, or they actually worsen inequality. Our work can be applied to a large range of applications (e.g., recruitment and bank lending). In these applications, an *institute* observes an *individual's* features (e.g., credit scores), and makes *myopic decisions* (e.g., issue loans) by assessing such features against a variable of interest (e.g., ability to repay) which is always unobservable to the institute when making decisions. The individual reacts to the decision by investing in effort to either improve or maintain its qualification in the next time step; these actions collectively then change the qualification rate of the population. In summary, our main contributions are:

1. We analyze the equilibrium of qualification rates in different groups under a general class of fairness constraints (Section 3). We use a Partially Observed Markov Decision Process (POMDP) framework to model the sequential myopic decision making in different scenarios (Section 2). Using this model, we first prove that the optimal myopic policy is a threshold policy and provide a way to compute it. We then prove the existence of an equilibrium (in terms of long-term qualification rates) using this optimal policy, and provide sufficient conditions for a unique equilibrium.
2. We analyze the impact of fairness constraints on the disparity of qualification rates when the equilibrium is unique (Section 4). Our findings suggest that the same fairness constraint can have opposite impacts on the equilibrium depending on the underlying problem scenario.
3. We explore alternative interventions that can be effective in improving qualification rates at the equilibrium and promoting equality across different groups (Section 5).
4. We examine our theory on synthetic Gaussian datasets and two real-world scenarios (Section 6). Our experiments show consistent conclusions cross domains and support for real-life policy making.

2 Problem Formulation

Partially Observed Markov Decision Process (POMDP).

Consider two groups \mathcal{G}_a and \mathcal{G}_b , distinguished by a sensitive attribute $S = s \in \{a, b\}$ (e.g., gender), each of a fraction $p_s := \mathbb{P}(S = s)$ of the population. At time t , an *individual* with attribute s has features $X_t = x \in \mathbb{R}^d$ determined by a hidden *qualification state* $Y_t = y \in \{0, 1\}$, both time-varying. We adopt the natural assumption that an individual's attribute and current features constitute sufficient statistics, so that conditioned on these, the decision is independent of past features or decisions. This allows an *institute* (decision maker) to adopt a Markov policy: it makes decisions $D_t = d \in \{0, 1\}$ (reject or accept) using a policy $\pi_t^s(x) := \mathbb{P}(D_t = 1 \mid X_t = x, S = s)$, $x \in \mathbb{R}^d$, $s \in \{a, b\}$ to maximize an instantaneous utility $R_t(D_t, Y_t)$, possibly subject to certain constraints. An individual is informed of the decision, and subsequently takes actions that may change their qualification Y_{t+1} and features X_{t+1} . The latter of which is used to drive the institute's decision at the next time step. This process is shown in Fig. 1. Note that this model can be viewed as capturing either a randomly selected individual repeatedly going through the decision cycles, or population-wide average when all individuals are subject to the decision cycles. Thus, $\alpha_t^s := \mathbb{P}(Y_t = 1 \mid S = s)$ is the probability of an individual from \mathcal{G}_s qualified at time t at the individual level, while being the *qualification rate* at the group level. One of our primary goals is to study how (α_t^a, α_t^b) evolves over time under different (fair) policies.

Feature generation process. At time step t , given $Y_t = y$ and $S = s$, features X_t are generated by $G_y^s(x) := \mathbb{P}(X_t = x \mid Y_t = y, S = s)$, $\forall x \in \mathbb{R}^d, y \in \{0, 1\}, s \in \{a, b\}$. This will be referred to as the *feature distribution* and assumed time-invariant. The convex combination $\mathbb{P}(X_t = x \mid S = s) = \alpha_t^s G_1^s(x) + (1 - \alpha_t^s) G_0^s(x)$ will be referred to as the *composite feature distribution* of group \mathcal{G}_s at time t . In Appendix G we consider a further generalization where the feature distribution is also affected by the previous decision: $G_{y,d}^s(x) := \mathbb{P}(X_t = x \mid Y_t = y, D_{t-1} = d, S = s)$.

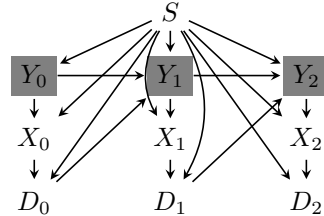


Figure 1: The graphical representation of our model where gray shades indicate latent variables.

Transition of qualification states. At time t , after receiving decision D_t , an individual takes actions such as exerting effort/investment, imitating others, etc., which results in a new qualification Y_{t+1} . This is modeled by a set of transitions $T_{yd}^s := \mathbb{P}(Y_{t+1} = 1 \mid Y_t = y, D_t = d, S = s), y \in \{0, 1\}, d \in \{0, 1\}, s \in \{a, b\}$. These transitions characterize individuals' ability to maintain or improve its qualification. Note that we don't model individuals' strategic responses as in [22, 29], but rather use T_{yd}^s to capture the overall effect; in other words, this single quantity may encapsulate the individual's willingness to exert effort, the cost of such effort, as well as the strength of community support, etc. Specifically, T_{0d}^s (resp. T_{1d}^s) represents the probability of individuals from \mathcal{G}_s who were previously unqualified (resp. qualified) became (resp. remain) qualified after receiving decision $d \in \{0, 1\}$.

Fair myopic policy of an institute. A myopic policy π_t at time t aims at maximizing the instantaneous expected utility/reward $\mathcal{U}(D_t, Y_t) = \mathbb{E}[R_t(D_t, Y_t)]$, where the institute gains $u_+ > 0$ by accepting a qualified individual and incurs a cost $u_- > 0$ by accepting an unqualified individual,

$$\text{i.e., } R_t(D_t, Y_t) := \begin{cases} u_+, & \text{if } Y_t = 1 \text{ and } D_t = 1 \\ -u_-, & \text{if } Y_t = 0 \text{ and } D_t = 1. \\ 0, & \text{if } D_t = 0 \end{cases}$$

A fair myopic policy maximizes the above utility subject to a fairness constraint \mathcal{C} . We focus on a set of group fairness constraints that equalize certain statistical measure between \mathcal{G}_a and \mathcal{G}_b . A commonly studied (one-shot) fair machine learning problem is to find (π_t^a, π_t^b) that solves the following constrained optimization,

$$\begin{aligned} \max_{\pi_t^a, \pi_t^b} \quad & \mathcal{U}(D_t, Y_t) = p_a \mathbb{E}[R_t(D_t, Y_t) \mid S = a] + p_b \mathbb{E}[R_t(D_t, Y_t) \mid S = b] \\ \text{s.t.} \quad & \mathbb{E}_{X_t \sim \mathcal{P}_C^a}[\pi_t^a(X_t)] = \mathbb{E}_{X_t \sim \mathcal{P}_C^b}[\pi_t^b(X_t)], \end{aligned} \quad (1)$$

where \mathcal{P}_C^s is some probability distribution over features X_t and specifies the fairness metric \mathcal{C} . Many popular fairness metrics can be written in this form, e.g.,

1. Equal of Opportunity (EqOpt) [17]: this requires the true positive rate (TPR) be equal, i.e., $\mathbb{P}(D_t = 1 \mid Y_t = 1, S = a) = \mathbb{P}(D_t = 1 \mid Y_t = 1, S = b)$. This is equivalent to $\mathbb{E}_{X_t \mid Y_t=1, S=a}[\pi_t^a(X_t)] = \mathbb{E}_{X_t \mid Y_t=1, S=b}[\pi_t^b(X_t)]$, i.e., $\mathcal{P}_{\text{EqOpt}}^s(x) = G_1^s(x)$.
2. Demographic Parity (DP) [5]: this requires the positive rate (PR) be equal, i.e., $\mathbb{P}(D_t = 1 \mid S = a) = \mathbb{P}(D_t = 1 \mid S = b)$. This is equivalent to $\mathbb{E}_{X_t \mid S=a}[\pi_t^a(X_t)] = \mathbb{E}_{X_t \mid S=b}[\pi_t^b(X_t)]$, i.e., $\mathcal{P}_{\text{DP}}^s(x) = (1 - \alpha_t^s)G_0^s(x) + \alpha_t^s G_1^s(x)$.

We focus on this class of myopic policies in this paper, and refer to the solution to (1) as the optimal policy. We further define *qualification profile*¹, $\gamma_t^s(x)$, the probability an individual with features x from group \mathcal{G}_s is qualified at t , i.e.,

$$\gamma_t^s(x) = \mathbb{P}(Y_t = 1 \mid X_t = x, S = s) = \frac{1}{\frac{G_0^s(x)}{G_1^s(x)}(\frac{1}{\alpha_t^s} - 1) + 1}, \quad x \in \mathbb{R}^d. \quad (2)$$

Then the utility obtained from the group \mathcal{G}_s at time step t is given by $\mathbb{E}[R_t(D_t, Y_t) \mid S = s] = \mathbb{E}_{X_t \mid S=s}[\pi_t^s(X_t)(\gamma_t^s(X_t)(u_+ + u_-) - u_-)]$. See Appendix D for details on the derivation.

3 Evolution and Equilibrium Analysis of Qualification Rates

In this section we first solve the one-shot optimization problem (1) (Sec. 3.1). We then show that under the optimal policy, there exists an equilibrium of qualification rates in the long run; a sufficient condition for its uniqueness is also introduced (Sec. 3.2). For simplicity of exposition, our analysis from here on out is for one-dimensional feature space; our conclusions, however, hold for high-dimensional features; this can be done by mapping features to the one-dimensional qualification profile space which admits optimal threshold policies (this extension is given in Appendix E).

3.1 Threshold policies are optimal

If an individual's qualification is observable, the optimal policy is straightforward absent of fairness constraints: accepting all qualified ones and rejecting the rest. When qualification is not observable, the institute needs to infer from the observed feature and accepts those most likely to be qualified. Next we show that under mild assumptions, optimal policies are in the form of threshold policies.

¹We assume the institute has perfect knowledge of $\gamma_t^s(x)$. In practice, this is obtained via learning/estimating α_t^s and $G_y^s(x)$ from data [24, 39]. The impact of estimation error is discussed in the Appendix F.

Assumption 1. $\forall s \in \{a, b\}$, $G_y^s(x)$ is continuous and $\frac{G_0^s(x)}{G_1^s(x)}$ is strictly decreasing in $x \in \mathbb{R}$; and the CDF $\mathbb{G}_y^s(x) = \int_{-\infty}^x G_y^s(z)dz$ is continuous in $x \in \mathbb{R}$, $\forall y \in \{0, 1\}$.

Assumption 2. $\forall s \in \{a, b\}$, $\mathcal{P}_C^s(x)$ is continuous in $x \in \mathbb{R}$; $\frac{G_1^s(x)}{\mathcal{P}_C^s(x)}$ is non-decreasing in $x \in \mathbb{R}$; and $\frac{G_0^s(x)}{\mathcal{P}_C^s(x)}$ is non-increasing in $x \in \mathbb{R}$.

Assumption 1 says that an individual is more likely to be qualified as his or her feature value increases². Assumption 2 limits the type of fairness constraints, but is satisfied by many commonly used ones, including EqOpt and DP.

Theorem 1 (Threshold policies are optimal). *Under Assumption 1, the unconstrained optimal policy is a threshold policy; for any fairness constraint \mathcal{C} satisfying Assumption 2, the optimal fair policy is a threshold policy, i.e., $\forall x, t$ and $s \in \{a, b\}$, $\pi_t^s(x) = \mathbf{1}(x \geq \theta_t^s)$ for some $\theta_t^s \in \mathbb{R}$. Threshold as a function of qualification rates, $\theta_t^s := \theta_t^s(\alpha_t^a, \alpha_t^b)$, is continuous and non-increasing in α_t^a and α_t^b .*

Theorem 1 shows that under mild assumptions, the optimal (fair) policies are deterministic threshold policies. We characterize these optimal (fair) thresholds in the next Lemma 1.

Lemma 1 (Optimal (fair) threshold). *Let $(\gamma^a(x), \gamma^b(x))$ be a pair of qualification profiles for groups \mathcal{G}_a and \mathcal{G}_b at t . Let threshold pairs $(\theta_{\mathcal{U}\mathcal{N}}^{a*}, \theta_{\mathcal{U}\mathcal{N}}^{b*})$ and $(\theta_{\mathcal{C}}^{a*}, \theta_{\mathcal{C}}^{b*})$ be the unconstrained and fair optimal thresholds under constraint \mathcal{C} , respectively. Then we have $\gamma^a(\theta_{\mathcal{U}\mathcal{N}}^{a*}) = \gamma^b(\theta_{\mathcal{U}\mathcal{N}}^{b*}) = \frac{u_-}{u_+ + u_-}$ and*

$$p_a \left(\gamma^a(\theta_{\mathcal{C}}^{a*}) - \frac{u_-}{u_+ + u_-} \right) \frac{\mathbb{P}(X = \theta_{\mathcal{C}}^{a*} | S = a)}{\mathcal{P}_C^a(\theta_{\mathcal{C}}^{a*})} + p_b \left(\gamma^b(\theta_{\mathcal{C}}^{b*}) - \frac{u_-}{u_+ + u_-} \right) \frac{\mathbb{P}(X = \theta_{\mathcal{C}}^{b*} | S = b)}{\mathcal{P}_C^b(\theta_{\mathcal{C}}^{b*})} = 0. \quad (3)$$

Here we have removed the subscript t since the thresholds are not t -dependent; they only depend on current qualification rates. The solution to Eqn. (3) is the threshold pair $(\theta_{\mathcal{C}}^{a*}, \theta_{\mathcal{C}}^{b*})$ that satisfies the fairness constraint $\int_{\theta_{\mathcal{C}}^{a*}}^{\infty} \mathcal{P}_C^a(x)dx = \int_{\theta_{\mathcal{C}}^{b*}}^{\infty} \mathcal{P}_C^b(x)dx$ in Eqn. (1) while maximizing the expected utility $\mathcal{U}(D, Y)$ at time t . Under DP and EqOpt fairness, Eqn. (3) can be reduced to

$$p_a \gamma^a(\theta_{\text{DP}}^{a*}) + p_b \gamma^b(\theta_{\text{DP}}^{b*}) = \frac{u_-}{u_+ + u_-}; \quad \frac{p_a \alpha^a}{\gamma^a(\theta_{\text{EqOpt}}^{a*})} + \frac{p_b \alpha^b}{\gamma^b(\theta_{\text{EqOpt}}^{b*})} = \frac{p_a \alpha^a}{\frac{u_-}{u_+ + u_-}} + \frac{p_b \alpha^b}{\frac{u_-}{u_+ + u_-}}.$$

Lemma 1 also indicates the relation between the unconstrained and fair optimal policies, e.g., a group's qualification profile evaluated at the unconstrained threshold is the same as the weighted combination of two groups' qualification profiles evaluated at their corresponding fair thresholds under DP.

3.2 Evolution and equilibrium analysis

We next examine what happens as the institute repeatedly makes decisions based on the optimal (fair) policies derived in Sec. 3.1, while individuals react by taking actions to affect their future qualifications. We say the qualification rate of \mathcal{G}_s is at an *equilibrium* if $\alpha_{t+1}^s = \alpha_t^s, \forall t \geq t_o$ for some t_o , or equivalently, if $\lim_{t \rightarrow \infty} \alpha_t^s = \alpha^s$ is well-defined for some $\alpha^s \in [0, 1]$. Analyzing equilibrium helps us understand the property of the population in the long-run. We begin by characterizing the dynamics of qualification rates α_t^s under policy π_t^s as follows (see Appendix D for the derivation):

$$\alpha_{t+1}^s = g_t^{0s}(1 - \alpha_t^s) + g_t^{1s}\alpha_t^s, \quad s \in \{a, b\}, \quad (4)$$

where $g_t^{ys} := \mathbb{E}_{X_t | Y_t = y, S = s} \left[(1 - \pi_t^s(X_t)) T_{y0}^s + \pi_t^s(X_t) T_{y1}^s \right]$. When $\pi_t^s(x) = \mathbf{1}(x \geq \theta_t^s)$, this reduces to $g_t^{ys} := \int_{-\infty}^{\theta_t^s} T_{y0}^s G_y^s(x)dx + \int_{\theta_t^s}^{\infty} T_{y1}^s G_y^s(x)dx = T_{y0}^s \mathbb{G}_y^s(\theta_t^s) + T_{y1}^s (1 - \mathbb{G}_y^s(\theta_t^s)), y \in \{0, 1\}$.

Theorem 2 below shows that for any transition and any threshold policy that are continuous in qualification rates, the above dynamical system always has at least one equilibrium.

Theorem 2 (Existence of equilibrium). *Consider a dynamics (4) under a threshold policy $\theta^s(\alpha^a, \alpha^b)$ that is continuous in α^a and α^b . $\forall T_{yd} \in (0, 1)$, there exists at least one equilibrium $(\hat{\alpha}^a, \hat{\alpha}^b)$.*

While an equilibrium exists under any set of transitions, its specific property highly depends on the transitions $T_{yd}^s, y = 0, 1, d = 0, 1$. We focus on two scenarios given in the condition below.

²When qualification increases as the feature value x decreases, one can simply use the opposite of x .

Condition 1. $\forall s \in \{a, b\}$, (A) $T_{01}^s \leq T_{00}^s$ and $T_{11}^s \leq T_{10}^s$; (B) $T_{01}^s \geq T_{00}^s$ and $T_{11}^s \geq T_{10}^s$.

As mentioned earlier, transitions T_{yd}^s characterize the ability of individuals from \mathcal{G}_s to maintain/improve their future qualifications. On one hand, an accepted individual may feel less motivated to remain qualified (if it was) or become qualified (if it was not); on the other hand, the accepted individual may have access to better resources or feel more inspired to remain or become qualified. These competing factors (referred to later as the ‘‘motivation’’ effect and the ‘‘leg-up’’ effect, respectively) may work simultaneously, and the net effect can be context dependent. Condition 1(A) (resp. Condition 1(B)) suggests that the first (resp. second) effect is dominant for both qualified and unqualified individuals. There are two other combinations: (C) $T_{01}^s \geq T_{00}^s$ and $T_{11}^s \leq T_{10}^s$; (D) $T_{01}^s \leq T_{00}^s$ and $T_{11}^s \geq T_{10}^s$. These cases incur more uncertainty; slight changes in feature distributions or transitions may result in opposite conclusions. More discussions are in Appendix F.

Given the existence of an equilibrium, Theorem 3 further introduces a sufficient condition for it to be unique. Based on the unique equilibrium, we can evaluate and compare policies (Sec. 4), and design effective interventions to promote long-term equality and/or the overall qualifications (Sec. 5).

Theorem 3 (Uniqueness of equilibrium). *Consider a dynamics system (4) under either unconstrained or fair optimal threshold policy. Let $h^s(\theta^s(\alpha^a, \alpha^b)) := \frac{1-g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))}$, $s \in \{a, b\}$. Under Assumptions 1 and 2, a sufficient condition for (4) to have a unique equilibrium is as follows, $\forall s \in \{a, b\}$:*

1. Under Condition 1(A), $|\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^{-s}}| < 1$, $\forall \alpha^s \in [0, 1]$, where $-s := \{a, b\} \setminus s$;
2. Under Condition 1(B), $|\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^{-s}}| < 1$ and $|\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^s}| < 1$, $\forall \alpha^a, \alpha^b \in [0, 1]$.

These sufficient conditions can further be satisfied if for the qualified ($y = 1$) and the unqualified ($y = 0$), transitions T_{y1}^s and T_{y0}^s are sufficiently close (see Corollary 1, Appendix H), i.e., policies have limited influence on the qualification dynamics. Moreover, it is worth noting that the uniqueness can also be guaranteed if the dynamics (4) satisfies L -Lipschitz condition with $L < 1$, which is relatively stronger than the condition in Theorem 3 (see the comparison in Appendix F).

Figure 2 illustrates trajectories of qualification rates (α_t^a, α_t^b) and the equilibrium for a Gaussian case under Condition 1(B) (see details in Appendix A). Let $g^{ys} := g^{ys}(\theta^s(\alpha^a, \alpha^b))$, the points (α^a, α^b) on the red and blue dashed curves satisfy $\alpha^b = g^{0b}(1 - \alpha^a) + g^{1b}\alpha^a$ and $\alpha^a = g^{0a}(1 - \alpha^a) + g^{1a}\alpha^a$, respectively. Their intersection (black star) is

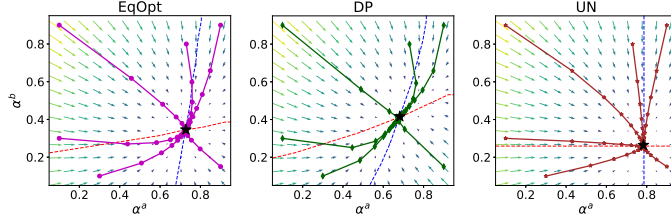


Figure 2: Illustration of $\{(\alpha_t^a, \alpha_t^b)\}_t$ for a Gaussian case under EqOpt, DP, UN optimal policies: each plot shows 6 sample paths with each circle/diamond/star representing one pair of (α_t^a, α_t^b) .

the equilibrium $(\hat{\alpha}^a, \hat{\alpha}^b)$. The sufficient condition on $\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^u}$ in Theorem 3 guarantees these two curves have only one intersection. Moreover, observe that these two curves split the plane $\{(\alpha^a, \alpha^b) : \alpha^a \in [0, 1], \alpha^b \in [0, 1]\}$ into four parts, which can be used for determining how (α_t^a, α_t^b) will change at t . For example, if (α_t^a, α_t^b) falls into the left side of the blue dashed curve, then $\alpha_{t+1}^a > \alpha_t^a$; if (α_t^a, α_t^b) falls into the lower side of the red dashed curve, then $\alpha_{t+1}^b > \alpha_t^b$.

4 The Long-term Impact of Fairness Constraints

In this section, we analyze the long-term impact of imposing fairness constraints on the equality of group qualification; we will do so in the presence of *natural equality* (and inequality) [36] where equitable equilibria are attained naturally without imposing additional constraints (in our context, this means attaining $\hat{\alpha}_{\text{UN}}^a = \hat{\alpha}_{\text{UN}}^b$ using unconstrained policies).

Although there may exist multiple equilibria, in this section we will assume the conditions in Theorem 3 hold and limit ourselves to the unique equilibrium cases under DP and EqOpt, thereby providing a theoretical foundation and an illustration of how their long-term impact can be compared. As shown below, these short-term fairness interventions may not necessarily promote long-term equity, and their impact can be sensitive to feature distributions and transitions. A small change in either can lead to contrarian results, suggesting the importance of understanding the underlying population.

Long-term impact on natural equality. When there is natural equality, an unconstrained optimal policy will result in two groups converging to the same qualification rate, thus rendering fairness constraints unnecessary. The interesting question here is whether applying a fairness constraint can disrupt the equality. The theorem below shows that the DP and EqOpt fairness will do harm if the feature distributions are different.

Theorem 4. For any feature distribution $\{G_y^s(x)\}_{s,y}, \forall s \in \{a, b\}, \forall y \in \{0, 1\}$, and $\forall \alpha_{UN} \in (0, 1)$, there exist transitions $\{T_{yd}^s\}_{y,d,s}$ satisfying either Condition 1(A) or Condition 1(B) such that $\hat{\alpha}_{UN}^a = \hat{\alpha}_{UN}^b = \alpha_{UN}$. In this case, if $G_y^a(x) \neq G_y^b(x)$ (resp. $G_y^a(x) = G_y^b(x)$), then imposing either $\mathcal{C} = DP$ or EqOpt fair optimal policies will violate (resp. maintain) equality, i.e., $\hat{\alpha}_C^a \neq \hat{\alpha}_C^b$ (resp. $\hat{\alpha}_C^a = \hat{\alpha}_C^b$).

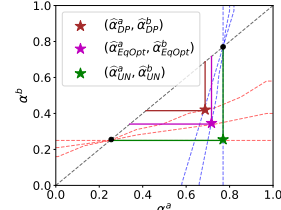
Theorem 4 shows that $\forall \alpha_{UN} \in (0, 1)$, there exists model parameters under which α_{UN} is the equilibrium and natural equality is attained. Also, natural equality is not disrupted by imposing either fairness constraint when feature distributions $G_y^s(x)$ are the same across different groups (referred to as *demographic-invariant* below). However, imposing either constraint will lead to unequal outcomes if $G_y^s(x)$ are diverse across groups (referred to as *demographic-variant* below), which is more likely to happen in reality. Thus, in these natural equality cases, imposing fairness will often do harm.

Long-term impact on natural inequality. Natural inequality, i.e., $\hat{\alpha}_{UN}^a \neq \hat{\alpha}_{UN}^b$, is more common than natural equality which only occurs under specific model parameters. This difference in qualification rates at equilibrium typically stems from the fact that either feature distributions or transitions are different across different groups. Below we examine the impact of imposing fairness by considering these two sources of inequality separately. Let *disadvantaged group* be the group with a lower qualification rate at equilibrium.

Demographic-invariant feature distribution with demographic-variant transition. In this case, we have feature distributions $G_y^a(x) = G_y^b(x), \forall x, y \in \{0, 1\}$, but different transitions. A real-world example is college admission based on ACT/SAT scores: given the same qualification state, score distributions may be the same regardless of the applicant's socio-economic status, but the economically advantaged may be able to afford more investments and effort to improve their score after a rejection.

Theorem 5. Under Condition 1(A), DP and EqOpt fairness exacerbate inequality, i.e., $|\hat{\alpha}_C^a - \hat{\alpha}_C^b| \geq |\hat{\alpha}_{UN}^a - \hat{\alpha}_{UN}^b|$; under Condition 1(B), DP and EqOpt fairness mitigate inequality, i.e., $|\hat{\alpha}_C^a - \hat{\alpha}_C^b| \leq |\hat{\alpha}_{UN}^a - \hat{\alpha}_{UN}^b|$. Moreover, the disadvantaged group remains disadvantaged in both cases, i.e., $(\hat{\alpha}_{UN}^a - \hat{\alpha}_{UN}^b)(\hat{\alpha}_C^a - \hat{\alpha}_C^b) \geq 0$.

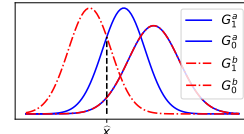
This theorem shows that imposing fairness only helps when the “leg-up” effect is more prominent than the “motivation” effect; alternatively, this suggests that when the “motivation” effect is dominant, imposing fairness should be accompanied by other support structure to dampen this effect (e.g., by helping those accepted to become or remain qualified). Theorem 5 is illustrated in the plot to the right, where transitions satisfy Condition 1(B) (mitigation happens) and $G_y^a(x) = G_y^b(x)$ is Gaussian distributed.



The plot includes 3 pairs of red/blue dashed curves corresponding to 3 policies (EqOpt, DP, UN). Points (α^a, α^b) on these curves satisfy $\alpha^b = g^{0b}(1 - \alpha^a) + g^{1b}\alpha^a$ and $\alpha^a = g^{0a}(1 - \alpha^a) + g^{1a}\alpha^a$, respectively. Each intersection (colored star) is an equilibrium $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$; the length of colored segments represents $|\hat{\alpha}_C^a - \hat{\alpha}_C^b|$. The black circle is the intersection of all three blue/red curves.

Demographic-variant feature distribution with demographic-invariant transition. In this case, we have transitions $T_{yd}^a = T_{yd}^b = T_{yd}, \forall y \in \{0, 1\}, \forall d \in \{0, 1\}$, but different feature distributions. In the same example of college admission this is a case where the ACT/SAT scores are biased against a certain group but there is no difference in how different groups react to the decision. Here, we will focus on a class of feature distributions where those qualified have the same feature distribution regardless of group membership, while those unqualified from \mathcal{G}_b are more likely to have lower features than those unqualified from \mathcal{G}_a . This is given in the condition below.

Condition 2. $G_y^s(x)$ is continuous in x ; $G_1^a(x) = G_1^b(x), \forall x$; $G_0^b(x) \geq G_0^a(x), \forall x$; \exists a unique \hat{x} such that $G_0^b(\hat{x}) = G_0^a(\hat{x}), G_0^b(x) > G_0^a(x), \forall x < \hat{x}$ and $G_0^b(x) < G_0^a(x), \forall x > \hat{x}$.



For bell-shaped distributions (e.g., Normal, Cauchy, etc.), Condition 2 can be satisfied if $G_0^b(x)$ is a mean-shifted version of $G_0^a(x)$. An example is shown on the right.

Theorem 6. *Under Condition 2 and Condition 1(B), if $\frac{u_+}{u_-} \geq \frac{G_0^s(\hat{x})}{G_1^s(\hat{x})} \frac{1-T_{10}}{T_{00}}$, we have*

- $\hat{\alpha}_{UN}^a > \hat{\alpha}_{UN}^b$ and $\hat{\alpha}_{UN}^a - \hat{\alpha}_{UN}^b > \hat{\alpha}_{EqOpt}^a - \hat{\alpha}_{EqOpt}^b \geq 0$ hold, i.e., *EqOpt fairness always mitigates inequality and the disadvantaged group \mathcal{G}_b remains disadvantaged.*
- *DP fairness may either (1) mitigate inequality, i.e., $\hat{\alpha}_{UN}^a - \hat{\alpha}_{UN}^b > \hat{\alpha}_{DP}^a - \hat{\alpha}_{DP}^b \geq 0$; or (2) flip the disadvantaged group from \mathcal{G}_b to \mathcal{G}_a , i.e., $\hat{\alpha}_{DP}^b \geq \hat{\alpha}_{DP}^a$.*

Because \mathcal{G}_a and \mathcal{G}_b only differ in $G_0^s(x)$, the condition in Thm 6 ensures at least one group has enough unqualified people to be accepted and can be satisfied if benefit u_+ is sufficiently larger than cost u_- . We see that in this case the comparison is much more complex depending on the model parameters.

5 Effective interventions

As discussed, imposing static fairness constraints is not always a valid intervention in terms of its long-term impact. In some cases it reinforces existing disparity; even when it could work, doing it right can be very hard due to its sensitivity to problem parameters. In this section, we present several alternative interventions that can be more effective in inducing more equitable outcomes or improving overall qualification rates in the long run. We shall assume that the sufficient conditions of Theorem 3 hold so that the equilibrium is unique in this section.

Policy intervention. In many instances, preserving static fairness at each time t is important, for short-term violations may result in costly lawsuits [1]. Proposition 1 below shows that sacrificing short-term utility while maintaining static fairness can improve overall qualification in the long run.

Proposition 1. *Let (θ_C^a, θ_C^b) , $(\theta_C^{a'}, \theta_C^{b'})$ be thresholds satisfying fairness constraint \mathcal{C} under the optimal and an alternative policy, respectively. Let $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$, $(\hat{\alpha}_C^{a'}, \hat{\alpha}_C^{b'})$ be the corresponding equilibrium.*

- *If $\theta_C^{s'}(\alpha^a, \alpha^b) > \theta_C^s(\alpha^a, \alpha^b)$, $\forall \alpha^s \in [0, 1]$ under Condition 1(A), then $\hat{\alpha}_C^{s'} > \hat{\alpha}_C^s$, $\forall s \in \{a, b\}$;*
- *If $\theta_C^{s'}(\alpha^a, \alpha^b) < \theta_C^s(\alpha^a, \alpha^b)$, $\forall \alpha^s \in [0, 1]$ under Condition 1(B), then $\hat{\alpha}_C^{s'} > \hat{\alpha}_C^s$, $\forall s \in \{a, b\}$.*

Note that the sacrifice is in instantaneous utility, not necessarily in total utility in the long run (see an example in proof of Proposition 1, Appendix H). If static fairness need not be maintained at all times, then we can employ separate policies for each group, and Proposition 2 below shows that under certain conditions on transitions, threshold policies leading to equitable equilibrium always exist.

Proposition 2. *Let $\mathcal{I}_s := \left[\frac{\min\{T_{01}^s, T_{00}^s\}}{1 - \min\{T_{11}^s, T_{10}^s\} + \min\{T_{01}^s, T_{00}^s\}}, \frac{\max\{T_{01}^s, T_{00}^s\}}{1 - \max\{T_{11}^s, T_{10}^s\} + \max\{T_{01}^s, T_{00}^s\}} \right]$, $s \in \{a, b\}$.*

Under Condition 1(A) or 1(B), if $\mathcal{I}_a \cap \mathcal{I}_b \neq \emptyset$, then $\forall \hat{\alpha} \in \mathcal{I}_a \cap \mathcal{I}_b$, there exist threshold policies $\theta^s(\alpha^s)$, $\forall \alpha^s \in [0, 1]$, under which $\alpha_t^s \rightarrow \hat{\alpha}$, $\forall s \in \{a, b\}$, i.e., equitable equilibrium is attained; if $\mathcal{I}_a \cap \mathcal{I}_b = \emptyset$, then there is no threshold policy that can result in equitable equilibrium.

Proposition 2 also indicates that when two groups' transitions are significantly different, manipulating policies cannot achieve equality. In this case, the following intervention can be considered.

Transition Intervention. Another intervention is to alter the value of transitions, e.g., by establishing support for both the accepted and rejected. Proposition 3 shows that the qualification rate $\hat{\alpha}^s$ at equilibrium can be improved by enhancing individuals' ability to maintain/improve qualification, which is consistent with the empirical findings in loan repayment [40, 16, 21] and labor markets [13].

Proposition 3. *$\forall s \in \{a, b\}$, increasing any transition probability T_{yd}^s , $d \in \{0, 1\}$, $y \in \{0, 1\}$ always increases the value of equilibrium qualification rates $\hat{\alpha}^s$.*

6 Experiments

We conducted experiments on both Gaussian synthetic datasets and real-world datasets. We present synthetic data experiments in Appendix A and the results using real-world datasets here. These are static, one-shot datasets, which we use to create a simulated dynamic setting as detailed below.

Loan repayment simulation study. We use the FICO score dataset [41] to study the long-term impact of fairness constraints EqOpt and DP and other interventions on loan repayment rates in the Caucasian group \mathcal{G}_C and the African American group \mathcal{G}_{AA} . With the pre-processed data in [17], we simulate a dataset with loan repayment records and credits scores. We first compute the initial qualification (loan repayment) rates and estimate the feature distributions $G_y^s(x)$ with beta

distributions based on the simulated data. Then, we compute the optimal UN, EqOpt, DP threshold according to Eqn. (3). Consequently, with the dynamics (4), we update the qualification rates in both groups. This process proceeds and qualification rates in both groups change over time.

Our results show consistent findings with studies in loan repayment literature [16, 40]. Specifically, [40] studied the loan repayment in group lending and pointed out that in practice effective training and leadership among the groups who were issued loans can increase their willingness to pay and improve the group repayment rate. Similar conclusion is also suggested by [16]. In our model, these interventions can be regarded as stimulating transitions (i.e., T_{y1}) to improve the future repayment rates. And the scenarios under such intervention would satisfy Condition 1(B). Fig. 3 shows that under Condition 1(B), increasing the transition T_{01}^s always increases qualification rates, and DP in general can result in a more equitable equilibrium than EqOpt. Fig. 3(a) shows that in Demographic-invariant (D-invariant) transition cases ($T_{yd}^{AA} = T_{yd}^C$): (1) \mathcal{G}_{AA} always remains as disadvantaged group; (2) when T_{10} is small, the inequality under UN optimal policies is small and the intervention on T_{01} only has minor effects on equality; when T_{10} is large (darker blue points), varying T_{01} can affect disparity significantly; (3) imposing DP attains equitable equilibria in general, which is robust to transitions and consistent with the conclusion in [36]; (4) when T_{10} is small, imposing EqOpt exacerbates inequality as T_{01} increases; while T_{10} is sufficient large, equality can be attained and robust to transitions. In Fig. 3(b), it shows that in D-variant transition cases, by setting $T_{yd}^{AA} < T_{yd}^C$, the inequality between \mathcal{G}_{AA} and \mathcal{G}_C further gets reinforced. In summary, the effectiveness of such intervention (increasing T_{01}) on promoting equality highly depends on the value of T_{10} and policies.

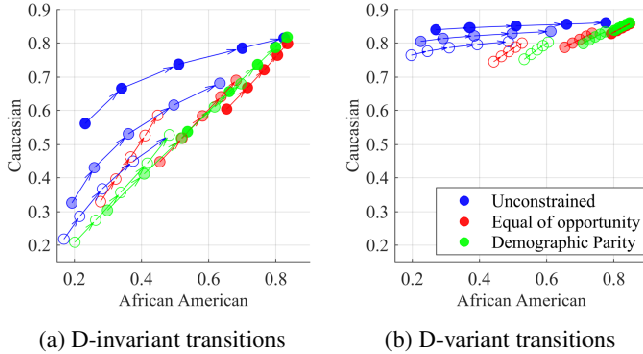


Figure 3: Results on the FICO dataset: Points are the equilibria of repayment rates in $\mathcal{G}_{AA}, \mathcal{G}_C$ under Condition 1(B) with different transitions. Arrows indicate the direction of increasing T_{01}^s ; a more transparent point represents the smaller value of T_{10}^s . In panel (a), $T_{yd}^{AA} = T_{yd}^C$, while in panel (b), $T_{yd}^{AA} < T_{yd}^C$.

The COMPAS data. Our second set of experiments is conducted on a multivariate recidivism prediction dataset from Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) [3]. We again use this static (and high-dimensional) dataset to create a simulated dynamic process as the FICO experiments. Specifically, from the raw data we calculate the initial qualification (recidivism) rate and train optimal classifier using a logistic regression model, based on which recidivism rate is updated according to Eqn. (4) under a given set of transitions. In the context of recidivism prediction, people who receive incarceration decisions are less likely to re-offend, thus we consider the transitions satisfying Condition 1(A). The classifier decision here corresponds to incarceration based on predicted likelihood of recidivism: the higher the predicted recidivism, the more likely an incarceration decision. In subsequent time steps, the data is re-sampled from the raw data proportional to the updated recidivism rates. This process repeats and the group recidivism rates change over time. Our results here primarily serve to highlight the complexity in such a dynamical system, and in particular, that an equilibrium may not exist.

Table 1 shows a typical result (details and other results in Appendix A) of four scenarios using an unconstrained policy and transitions under Condition 1 (A). In the first three cases/columns, the recidivism rate converges to an equilibrium. In particular, the first case/column uses the optimal threshold whereas the next two use a higher and a lower threshold compared to the optimum, respectively. We see that in the second column, a higher threshold (fewer incarcerations) can lead to higher recidivism, a phenomenon that has been observed in juvenile justice system history [35]. On the other hand, by changing the transitions, the fourth case (corresponding to the last two columns)

Table 1: Recidivism rates in the long run. UN*: unconstrained policy (UN) with the optimal threshold; UN_{θ_H} : UN with a higher threshold; UN_{θ_L} : UN with a lower threshold; UN_T : UN with different transitions, under which the system oscillates between two states.

	UN*	UN_{θ_H}	UN_{θ_L}	UN_T^1	UN_T^2
$\hat{\alpha}$	0.343	0.356	0.307	0.364	0.501

does not have an equilibrium and the system oscillates between the two states shown in the table, one much less desirable than the other (see more experiments for oscillation states in Appendix A).

7 Conclusion

In this paper, we study the long-term impact of fairness constraints (e.g., DP and EqOpt) on groups' qualification rates. In particular, we construct a POMDP for sequential decision making, and conduct equilibrium analysis based on it. Specifically, we first identify sufficient conditions for the existence and uniqueness of equilibrium, under which we compare different fairness constraints regarding their long-term impacts. Our findings show that the same fairness constraint can have opposite impacts depending on the underlying problem scenarios, which highlights the importance of understanding real-world dynamics in decision making systems. The experiments on real-world static datasets with simulated dynamics also show the consistent findings with social science studies. Our analysis has focused on scenarios with a unique equilibrium in this work. In the future, we will continue our work with a theoretical analysis of scenarios with multiple equilibria or oscillation states.

Broader Impact

In this paper, we focus on the (un)fairness issue that arises in automated decision-making systems and aim to understand the long-term impact of algorithmic (fair) decisions on the well-being of different sub-groups in a population. Our partially observed sequential decision making framework is applicable to a wide range of domains (e.g., lending, recruitment, admission, criminal justice, etc.). By conducting an equilibrium analysis and evaluating the long-term impact of different fairness criteria, our results provide a theoretical foundation that can help answer questions such as whether/when imposing short-term fairness constraints are effective in promoting long-term equality.

First of all, our results can help policymakers (e.g., companies, banks, governments, etc.) in their decision making process by highlighting the potential pitfalls of commonly used static fairness criteria and providing guidance on how to design effective interventions that can avoid such unintended consequences and result in positive long-term societal impacts.

Secondly, our results may be useful to research in fields outside of the computer science community. The experiments on static real-world datasets have shown consistent findings with literature in social sciences [35, 16, 40]. Although these empirical results are obtained using simulated dynamics due to a lack of real datasets, they may provide insights and theoretical supports for research in other fields.

Lastly, while this work is limited to binary decisions, the main take-away can be applied in other applications such as computer vision, natural language processing, etc., using more complicated classifiers such as DNN. We hope that our work will encourage researchers in these domains to similarly consider discrimination risks when developing techniques, and raise awareness that static fairness constraint may not suffice and long-term fairness cannot be designed in a vacuum without considering the human element. We thus emphasize the importance of performing real-time measurements and developing proper fair classifiers from dynamic datasets.

Having mentioned the potential positive impact of our work, we also want to point out the limitations in our model and analysis. Firstly, in this work we use a set of transitions T_{yd}^s to capture individuals' abilities to improve/maintain future qualifications, and our analysis and conclusions rely on this set of values. In practice, however, these quantities can be extremely hard to measure due to the complexity of human behaviors and environmental factors. In addition, as we have noted in the paper, in some cases the conclusion can be highly sensitive to minor changes in these transitions. Secondly, our theoretical results have focused on scenarios with a unique equilibrium, while in practice the situation can be much more complicated (multiple equilibria or no equilibria), as demonstrated by the oscillations we see in the COMPAS simulation study. Thus, it is worthwhile for future work to consider these more complex cases. Lastly, due to the lack of dynamic datasets, our experiments are performed over static real-world datasets with simulated dynamics. Thus, an accurate model of real-world dynamics is needed when deploying our method for practical decision making.

References

- [1] Target corporation to pay \$2.8 million to resolve eeoc discrimination finding. In *U.S. Equal Employment Opportunity Commission*, 2015. <https://www.eeoc.gov/newsroom/>

- target-corporation-pay-28-million-resolve-eeoc-discrimination-finding.
- [2] A. P. Aneja and C. F. Avenancio-León. No credit for time served? incarceration and credit-driven crime cycles. 2019.
 - [3] 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*. *ProPublica*, 23, 2016.
 - [4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.
 - [5] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. <http://www.fairmlbook.org>.
 - [6] Y. Bechavod, K. Ligett, A. Roth, B. Waggoner, and S. Z. Wu. Equal opportunity in online classification with partial feedback. In *Advances in Neural Information Processing Systems 32*, pages 8972–8982. 2019.
 - [7] A. J. Chaney, B. M. Stewart, and B. E. Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232. ACM, 2018.
 - [8] Y. Chen, A. Cuellar, H. Luo, J. Modi, H. Nemlekar, and S. Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. *arXiv preprint arXiv:1912.08055*, 2019.
 - [9] A. D’Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.
 - [10] C. Dimitrakakis, Y. Liu, D. Parkes, and G. Radanovic. Bayesian fairness. In *AAAI*, 2019.
 - [11] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference of Fairness, Accountability, and Transparency*, 2018.
 - [12] D. Ensign, F. Sorelle, N. Scott, S. Carlos, and V. Suresh. Decision making with limited feedback. In *Algorithmic Learning Theory*, pages 359–367, 2018.
 - [13] E. Fehr, L. Goette, and C. Zehnder. A behavioral account of the labor market: The role of fairness concerns. *Annu. Rev. Econ.*, 1(1):355–384, 2009.
 - [14] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets*, 2018.
 - [15] S. Gillen, C. Jung, M. Kearns, and A. Roth. Online learning with an unknown fairness metric. In *Advances in Neural Information Processing Systems*, pages 2600–2609, 2018.
 - [16] K. F. Gotham. Race, mortgage lending and loan rejections in a us city. *Sociological Focus*, 31(4):391–405, 1998.
 - [17] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
 - [18] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938. PMLR, 2018.
 - [19] H. Heidari and A. Krause. Preventing disparate treatment in sequential decision making. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2248–2254, 2018.
 - [20] H. Heidari, V. Nanda, and K. Gummadi. On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In *International Conference on Machine Learning*, pages 2692–2701, 2019.
 - [21] T. Homonoff, R. O’Brien, and A. B. Sussman. Does knowing your fico score change financial behavior? evidence from a field experiment with student loan borrowers. *Review of Economics and Statistics*, pages 1–45.

- [22] L. Hu, N. Immorlica, and J. W. Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.
- [23] S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, and A. Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1617–1626. JMLR. org, 2017.
- [24] I. James. Estimation of the mixing proportion in a mixture of two normal distributions from simple, rapid measurements. *Biometrics*, pages 265–275, 1978.
- [25] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.
- [26] M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163. ACM, 2018.
- [27] N. Kallus and A. Zhou. Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- [28] S. Kannan, A. Roth, and J. Ziani. Downstream effects of affirmative action. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 240–248. ACM, 2019.
- [29] M. Khajehnejad, B. Tabibian, B. Schölkopf, A. Singla, and M. Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.
- [30] N. Kilbertus, M. Gomez-Rodriguez, B. Schölkopf, K. Muandet, and I. Valera. Improving consequential decision making under imperfect predictions. *arXiv preprint arXiv:1902.02979*, 2019.
- [31] F. Li, J. Liu, and B. Ji. Combinatorial sleeping bandits with fairness constraints. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 1702–1710. IEEE, 2019.
- [32] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. Delayed impact of fair machine learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158. PMLR, 2018.
- [33] L. T. Liu, A. Wilson, N. Haghtalab, A. T. Kalai, C. Borgs, and J. Chayes. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 381–391, 2020.
- [34] Y. Liu, G. Radanovic, C. Dimitrakakis, D. Mandal, and D. C. Parkes. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875*, 2017.
- [35] C. A. Mallett. Disproportionate minority contact in juvenile justice: Today’s, and yesterdays, problems. *Criminal justice studies*, 31(3):230–248, 2018.
- [36] H. Mouzannar, M. I. Ohannessian, and N. Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, page 359–368, 2019.
- [37] R. Nabi, D. Malinsky, and I. Shpitser. Learning optimal fair policies. *Proceedings of machine learning research*, 97:4674, 2019.
- [38] V. Patil, G. Ghalme, V. Nair, and Y. Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *arXiv preprint arXiv:1907.10516*, 2019.
- [39] R. K. Patra and B. Sen. Estimation of a two-component mixture model with applications to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):869–893, 2016.
- [40] J. Paxton, D. Graham, and C. Thraen. Modeling group loan repayment behavior: New insights from burkina faso. *Economic Development and cultural change*, 48(3):639–655, 2000.
- [41] U. F. Reserve. Report to the congress on credit scoring and its effects on the availability and affordability of credit. *Board of Governors of the Federal Reserve System*, 2007.
- [42] W. Tang, C.-J. Ho, and Y. Liu. Fair bandit learning with delayed impact of actions. *arXiv preprint arXiv:2002.10316*, 2020.

- [43] M. Wen, O. Bastani, and U. Topcu. Fairness with dynamics. *arXiv preprint arXiv:1901.08568*, 2019.
- [44] J. Williams and J. Z. Kolter. Dynamic modeling and equilibria in fair decision making. *arXiv preprint arXiv:1911.06837*, 2019.
- [45] X. Zhang, M. M. Khalili, and M. Liu. Long-term impacts of fair machine learning. *Ergonomics in Design*, 2019. doi: 10.1177/1064804619884160.
- [46] X. Zhang, M. Khaliligarekani, C. Tekin, et al. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems*, pages 15243–15252, 2019.

A Additional results on experiments

Gaussian distributed synthetic data. We first verify the conclusions in Section 3 and 4 using the synthetic data, where $X_t | Y_t = y, S = s \sim \mathcal{N}(\mu_y^s, (\sigma^s)^2)$.

In Section 3, Figure 2 illustrates sample paths of $\{(\alpha_t^a, \alpha_t^b)\}_t$ under EqOpt, DP, UN optimal policies. The specific parameters are as follows: $[\mu_0^a, \mu_1^a, \mu_0^b, \mu_1^b] = [-5, 5, -5, 5]$, $[\sigma^a, \sigma^b] = [5, 5]$, $\frac{u_+}{u_-} = 1$, $p_a = p_b = 0.5$, $[T_{00}^a, T_{01}^a, T_{10}^a, T_{11}^a] = [0.4, 0.5, 0.5, 0.9]$, $[T_{00}^b, T_{01}^b, T_{10}^b, T_{11}^b] = [0.1, 0.5, 0.5, 0.7]$.

Table 2 and 3 illustrate the impacts of EqOpt and DP fairness on the equilibrium, where each column shows the value of $\hat{\alpha}_C^a - \hat{\alpha}_C^b$ when $C = \text{UN, EqOpt, DP}$ under different sets of parameters. Specifically, in Table 2, $p_a = p_b = 0.5$, $\frac{u_+}{u_-} = 1$, $[\mu_0^s, \mu_1^s, \sigma^s] = [-5, 5, 5], \forall s \in \{a, b\}$ and transitions satisfying either Condition 1(A) or 1(B) are randomly generated; in Table 3, transitions satisfying Condition 1(B) and $G_y^s(x)$ that satisfy Condition 2 are randomly generated, $\frac{u_+}{u_-}$ also satisfies the condition in Theorem 6. These results are consistent with Theorem 5 and 6.

Table 2: $\hat{\alpha}_C^a - \hat{\alpha}_C^b$ when $C = \text{UN, EqOpt, DP}$: $G_y^a(x) = G_y^b(x)$ and $T_{yd}^a \neq T_{yd}^b$.

	Condition 1(A)								
UN ($\times 10^{-2}$)	-18.45	16.89	19.82	-7.21	-16.34	-26.56	16.66	-6.03	-38.63
EqOpt ($\times 10^{-2}$)	-21.11	19.13	21.78	-7.62	-18.56	-29.21	18.14	-6.28	-41.52
DP ($\times 10^{-2}$)	-27.98	23.11	25.65	-8.90	-23.11	-33.22	21.09	-6.66	-43.35
	Condition 1(B)								
UN ($\times 10^{-2}$)	-19.05	18.18	-0.70	-58.80	-40.91	61.30	12.82	-44.67	2.66
EqOpt ($\times 10^{-2}$)	-18.40	17.98	-0.64	-57.62	-34.50	48.66	12.35	-41.43	2.61
DP ($\times 10^{-2}$)	-17.52	17.73	-0.57	-55.62	-28.97	36.10	11.69	-37.97	2.57

Table 3: $\hat{\alpha}_C^a - \hat{\alpha}_C^b$ when $C = \text{UN, EqOpt, DP}$: $G_y^a(x) \neq G_y^b(x)$ and $T_{yd}^a = T_{yd}^b$ under Condition 1(B).

UN ($\times 10^{-2}$)	1.88	26.35	2.12	0.38	5.64	12.35	11.70	0.20	4.12
EqOpt ($\times 10^{-2}$)	0.57	17.43	1.75	0.32	5.05	7.81	7.21	0.18	1.68
DP ($\times 10^{-4}$)	16.26	18.29	-5.94	-0.93	-2.25	1.47	0.92	-1.68	-0.80

FICO score data. From the pre-processed FICO dataset, we got $\mathbb{P}(X = x | S = s)$ and $\mathbb{P}(Y = 1 | X = x, S = s)$. In this experiment, we consider two demographic groups, 12% the African American \mathcal{G}_{AA} and 88% the Caucasian \mathcal{G}_C . According to the empirical feature distributions, we can first simulate the FICO dataset with credit scores X , repayment Y , and sensitive attribute S . We then compute the initial qualification (repayment) rates $(\alpha_0^{AA}, \alpha_0^C)$, which is 0.34 in \mathcal{G}_{AA} and 0.76 in \mathcal{G}_C ; and fit Beta distributions to get the feature distribution $\mathbb{P}(X = x | S = s, Y = y)$, as shown in Fig. 4. Since the feature distributions are the Beta distributions, we can compute optimal UN, EqOpt, DP thresholds directly using Eqn. (3) and update the repayment rates based on dynamics (4). This process proceeds and $(\alpha_t^{AA}, \alpha_t^C)$ changes over time.

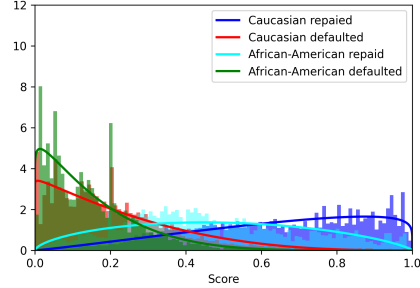


Figure 4: The feature distributions: the scores are rescaled so that they are between 0 and 1.

We then consider the demographic-invariant (D-invariant) and demographic-variant (D-variant) transitions and examine the impact of the transition interventions. Specifically, in the context of loan repayment prediction and group lending [40], the transitions would satisfy Condition 1(B). Fig. 5 illustrates the equilibria $(\hat{\alpha}^{AA}, \hat{\alpha}^C)$ under different sets of transitions. Their specific values are listed as follows, where the system has an equilibrium in all cases.

$$\begin{aligned}
 \text{D-invariant:} \quad & T_{00} = 0.1, T_{11} = 0.9, \quad T_{10}, T_{01} \in \{0.1, 0.3, 0.5, 0.7, 0.9\} \\
 \text{D-variant:} \quad & T_{00}^{AA} = 0.1, T_{11}^{AA} = 0.9, \quad T_{10}^{AA}, T_{01}^{AA} \in \{0.20, 0.36, 0.53, 0.69, 0.85\} \\
 & T_{00}^C = 0.4, T_{11}^C = 0.9, \quad T_{10}^C, T_{01}^C \in \{0.45, 0.55, 0.65, 0.75, 0.85\}
 \end{aligned}$$

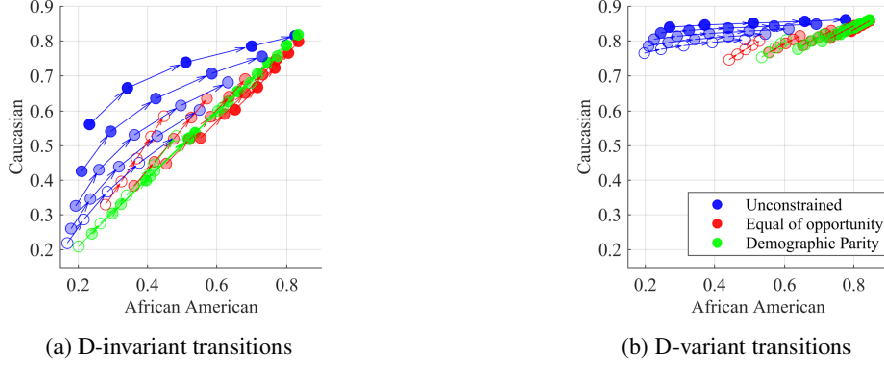


Figure 5: Results on the FICO dataset: Points are the repayment rates of \mathcal{G}_{AA} , \mathcal{G}_C at the equilibria under Condition 1(B) with different sets of transitions. Arrows indicate the direction of increasing T_{01}^s ; a more transparent point represents the smaller value of T_{10}^s . In panel (a), $T_{yd}^{AA} = T_{yd}^C$, while in panel (b), $T_{yd}^{AA} < T_{yd}^C$.

COMPAS data. The COMPAS dataset is a high-dimensional dataset with mixed data types (e.g., continuous, binary, and categorical). The number of samples is 5278. There are 10 features and two demographic groups: 60% African American (\mathcal{G}_{AA}) and 40% Caucasian (\mathcal{G}_C). The qualification rate in COMPAS is the recidivism rate. The initial recidivism rates are 52.3% in \mathcal{G}_{AA} and 39.1% in \mathcal{G}_C .

Due to the complexity of the feature distribution, the system can be either in the equilibrium state or oscillate between two recidivism rates in the long-run. Since the feature distribution is fixed and approximated from the COMPAS dataset, we investigate that under which transitions, the system is in an equilibrium state under unconstrained optimal policy. For this purpose, it is sufficient to study the demographic-invariant transitions $T^{AA} = T^C$ and consider the entire population without distinguishing two groups; moreover, in the context of recidivism prediction, the transitions would satisfy Condition 1 (A). Therefore, we consider T_{00} and T_{10} taking the values 0.1, 0.3, 0.5, 0.7 and 0.9. Figure 6 shows the results when $T_{01} = k \times T_{00}$ and $T_{11} = k \times T_{10}$. We find that when Corollary 1 is satisfied, e.g., when $k \geq 0.5$, most of the corresponding systems have a unique equilibrium (blue dot). Moreover, when $T_{00} \leq 0.5$, the system is also mostly in the unique equilibrium state. For the other transitions, the system oscillates between two states (red circle). We also show the results under all the combinations of T_{01} and T_{11} in Figure 7.

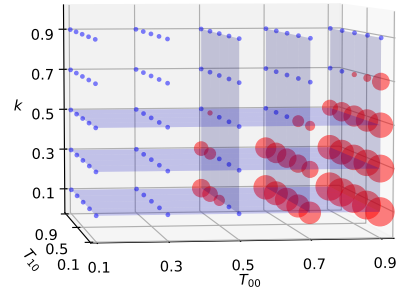


Figure 6: The oscillation level of recidivism rates in the long run is represented by the size of red circles, of which the bigger one represents severer oscillation. The blue dots represent the scenarios with a unique equilibrium. T_{00} and T_{10} axes represent their values respectively; k axis represents the scalar k , where $T_{01} = k \times T_{00}$ and $T_{11} = k \times T_{10}$.

Next, we study the impact of policy interventions in cases with equilibrium. We randomly choose the transitions under which the system has an equilibrium and then apply the unconstrained policy with optimal threshold (classifier threshold 0.5), a higher and a lower threshold (classifier thresholds 0.8 and 0.2 respectively) compared to the optimum respectively. The results are show in Table 4.

Table 4: Recidivism rates in the long run. UN*: unconstrained policy (UN) with the optimal threshold; UN_{θ_H} : UN with a higher threshold; UN_{θ_L} : UN with a lower threshold.

	UN*	UN_{θ_H}	UN_{θ_L}
$\hat{\alpha}_1$	0.164	0.166	0.147
$\hat{\alpha}_2$	0.343	0.356	0.307
$\hat{\alpha}_3$	0.230	0.246	0.162
$\hat{\alpha}_4$	0.306	0.3415	0.156
$\hat{\alpha}_5$	0.162	0.166	0.140

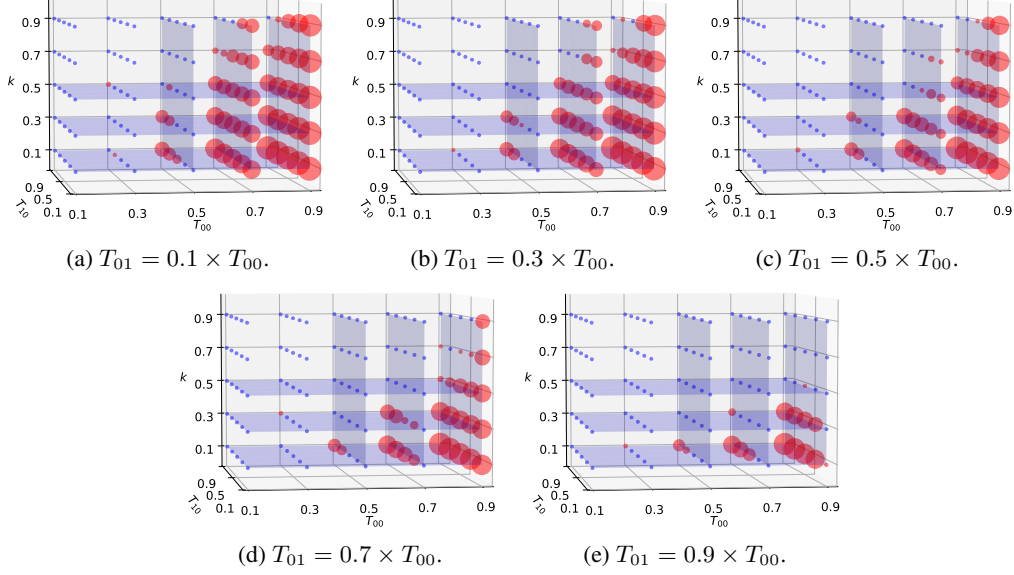


Figure 7: The oscillation level of recidivism rates under different transitions. In each panel, scalar k denotes the ratio, of which $T_{11} = k \times T_{10}$.

B Notations

\mathcal{G}_s	demographic group, $s \in \{a, b\}$
X_t	feature at t , $X \in \mathbb{R}^d$
Y_t	true qualification state at t , $Y_t \in \{0, 1\}$
S	sensitive attribute $S \in \{a, b\}$
p_s	group proportion of s , i.e., $p_s = \mathbb{P}(S = s)$
D_t	institute's decision at t , $D_t \in \{0, 1\}$
$\pi_t^s(x)$	policy for \mathcal{G}_s at t , i.e., $\pi_t^s(x) = \mathbb{P}(D_t = 1 \mid X_t = x, S = s)$
$G_y^s(x)$	feature distribution of unqualified ($y = 0$) or qualified ($y = 1$) people from \mathcal{G}_s , i.e., $\mathbb{P}(X_t = x \mid Y_t = y, S = s)$
$\mathbb{G}_y^s(x)$	CDF of $G_y^s(x)$, i.e., $\mathbb{G}_y^s(x) = \int_{-\infty}^x G_y^s(z) dx$
$\mathcal{P}_C^s(x)$	a probability distribution over X_t that specifies the fairness metric \mathcal{C}
α_t^s	qualification rate of \mathcal{G}_s at t , i.e., $\mathbb{P}(Y_t = 1 \mid S = s)$
$\gamma_t^s(x)$	qualification profile of \mathcal{G}_s at t , i.e., $\mathbb{P}(Y_t = 1 \mid X_t = x, S = s)$
T_{yd}^s	transition probability of \mathcal{G}_s , i.e., $\mathbb{P}(Y_{t+1} = 1 \mid Y_t = y, D_t = d, S = s)$
u_+	benefit the institute gains by accepting a qualified individual
u_-	cost incurred to the institute by accepting an unqualified individual
θ_C^s	threshold in a threshold policy for \mathcal{G}_s under constraint \mathcal{C} , i.e., $\pi_t^s(x) = \mathbf{1}(x \geq \theta_C^s)$
$\hat{\alpha}_C^s$	qualification rate of \mathcal{G}_s at the equilibrium under policy with constraint $\mathcal{C} \in \{\text{UN}, \text{DP}, \text{EqOpt}\}$

C Related Work

Among existing works on fairness in sequential decision making problems, many assume that the population's feature distribution neither changes over time nor is it affected by decisions; examples include studies on handling bias in online learning [6, 10–12, 15, 19, 27, 30] and bandits problems [4, 8, 25, 26, 31, 34, 38, 42]. The goal of most of these work is to design algorithms that can learn near-optimal policy quickly from the sequentially arrived data and the partially observed information, and understand the impact of imposing fairness intervention on the learned policy (e.g., total utility, learning rate, sample complexity, etc.)

However, recent studies [2, 7, 14] have shown that there exists a complex interplay between algorithmic decisions and individuals, e.g., user participation dynamics [18, 45, 46], strategic reasoning in

a game [22, 29], etc., such that decision making directly leads to changes in the underlying feature distribution, which then feeds back into the decision making process. Many studies thus aim at understanding the impacts of imposing fairness constraints when decisions affect underlying feature distribution. For example, [32, 20, 28, 29] construct two-stage models where only the one-step impacts of fairness intervention on the underlying population are examined but not the long-term impacts in a sequential framework; [23, 37, 43] focus on the fairness in reinforcement learning, of which the goal is to learn a long-run optimal policy that maximizes the cumulative rewards subject to certain fairness constraint; [18, 46] construct a user participation dynamics model where individuals respond to perceived decisions by leaving the system uniformly at random. The goal is to understand the impact of various fairness interventions on group representation.

Our work is most relevant to [22, 33, 36, 44], which study the long-term impacts of decisions on the groups' qualification states with considering dynamics. In [22, 33], strategic individuals are assumed to be able to observe the current policy, based on which they can manipulate the qualification states strategically to receive better decisions. However, there is a lack of study on the influence of the sensitive attribute on dynamics and impact of fairness constraints. Besides, in many cases, the qualification states are affected by both the policy and the qualifications at the previous time step, which is considered in [36, 44]. However, they assume that the decision maker can fully get access to qualification states and the dynamics of the qualification rates is the same in different demographic groups, i.e., they assume the equally qualified people from different groups after perceiving the same decision will have the same future qualification state. In fact, the qualification states are unobservable in most cases, and the dynamics can vary across different groups. If considering such difference, the dynamics can be much more complicated such that the social equality can not be attained easily as concluded in [36, 44].

D Derivations

Qualification profile of a group.

$$\begin{aligned}
\gamma_t^s(x) &= \mathbb{P}(Y_t = 1 | X_t = x, S = s) = \frac{1}{\frac{\mathbb{P}(X_t=x, Y_t=0, S=s)}{\mathbb{P}(X_t=x, Y_t=1, S=s)} + 1} \\
&= \frac{1}{\frac{\mathbb{P}(X_t=x | Y_t=0, S=s) \mathbb{P}(Y_t=0 | S=s)}{\mathbb{P}(X_t=x | Y_t=1, S=s) \mathbb{P}(Y_t=1 | S=s)} + 1} \\
&= \frac{1}{\frac{\mathbb{P}(X_t=x | Y_t=0, S=s)}{\mathbb{P}(X_t=x | Y_t=1, S=s)} \left(\frac{1}{\mathbb{P}(Y_t=1 | S=s)} - 1 \right) + 1} \\
&= \frac{1}{\frac{G_0^s(x)}{G_1^s(x)} \left(\frac{1}{\alpha_t^s} - 1 \right) + 1}.
\end{aligned}$$

Utility of an institute.

$$\mathcal{U}(D_t, Y_t) = \mathbb{E}[R_t(D_t, Y_t)] = \mathbb{P}(S = a) \mathbb{E}[R_t(D_t, Y_t) | S = a] + \mathbb{P}(S = b) \mathbb{E}[R_t(D_t, Y_t) | S = b]$$

Under policy π^s , we have

$$\begin{aligned}
&\mathbb{E}[R_t(D_t, Y_t) | S = s] = \mathbb{P}(D_t = 1, Y_t = 1 | S = s) u_+ - \mathbb{P}(D_t = 1, Y_t = 0 | S = s) u_- \\
&= \int_x \left(\mathbb{P}(D_t = 1, Y_t = 1, X_t = x | S = s) u_+ - \mathbb{P}(D_t = 1, Y_t = 0, X_t = x | S = s) u_- \right) dx \\
&= \int_x \mathbb{P}(X_t = x | S = s) \left(\mathbb{P}(D_t = 1 | X_t = x, S = s) \mathbb{P}(Y_t = 1 | X_t = x, S = s) u_+ \right. \\
&\quad \left. - \mathbb{P}(D_t = 1 | X_t = x, S = s) \mathbb{P}(Y_t = 0 | X_t = x, S = s) u_- \right) dx \\
&= \int_x \mathbb{P}(X_t = x | S = s) \left(\pi^s(x) \gamma_t^s(x) u_+ - \pi^s(x) (1 - \gamma_t^s(x)) u_- \right) dx \\
&= \mathbb{E}_{X_t | S=s} [\pi^s(X_t) (\gamma_t^s(X_t) (u_+ + u_-) - u_-)].
\end{aligned}$$

Therefore,

$$\mathcal{U}(D_t, Y_t) = p_a \mathbb{E}_{X_t | S=a} [\pi^a(X_t) (\gamma_t^a(X_t) (u_+ + u_-) - u_-)] + p_b \mathbb{E}_{X_t | S=b} [\pi^b(X_t) (\gamma_t^b(X_t) (u_+ + u_-) - u_-)]$$

Dynamics of qualification rate.

$$\begin{aligned}
\alpha_{t+1}^s &= \mathbb{P}(Y_{t+1} = 1 \mid S = s) = \int_x \sum_{y,a} \mathbb{P}(Y_{t+1} = 1, Y_t = y, D_t = d, X_t = x \mid S = s) dx \\
&= \int_x \sum_{y,a} \mathbb{P}(Y_{t+1} = 1 \mid Y_t = y, X_t = x, D_t = d, S = s) \mathbb{P}(D_t = d \mid X_t = x, S = s) \\
&\quad \mathbb{P}(X_t = x \mid Y_t = y, S = s) \mathbb{P}(Y_t = y \mid S = s) dx \\
&= \int_x \sum_a \left\{ \mathbb{P}(Y_{t+1} = 1 \mid Y_t = 0, X_t = x, D_t = d, S = s) \right. \\
&\quad \left. \mathbb{P}(D_t = d \mid X_t = x, S = s) \mathbb{P}(X_t = x \mid Y_t = 0, S = s) \right\} \mathbb{P}(Y_t = 0 \mid S = s) dx \\
&\quad + \int_x \sum_d \left\{ \mathbb{P}(Y_{t+1} = 1 \mid Y_t = 1, X_t = x, D_t = d, S = s) \right. \\
&\quad \left. \mathbb{P}(D_t = d \mid X_t = x, S = s) \mathbb{P}(X_t = x \mid Y_t = 1, S = s) \right\} \mathbb{P}(Y_t = 1 \mid S = s) dx \\
&= \mathbb{E}_{X_t \mid Y_t=0, S=s} \left[(1 - \pi_t^s(X_t)) T_{00}^s + \pi_t^s(X_t) T_{01}^s \right] (1 - \alpha_t^s) \\
&\quad + \mathbb{E}_{X_t \mid Y_t=1, S=s} \left[(1 - \pi_t^s(X_t)) T_{10}^s + \pi_t^s(X_t) T_{11}^s \right] \alpha_t^s \\
&= g_t^{0s} (1 - \alpha_t^s) + g_t^{1s} \alpha_t^s
\end{aligned}$$

E Generalization to high-dimensional feature space

All analysis and conclusions in this paper can be generalized to high-dimensional feature space $x \in \mathbb{R}^d$, where the qualification profile of \mathcal{G}_s is defined as $\gamma_t^s(\mathbf{x}) = \mathbb{P}(Y_t = 1 \mid \mathbf{X}_t = \mathbf{x}, S = s) \in [0, 1]$, $\mathbf{x} \in \mathbb{R}^d$. Different from one-dimensional case where decisions are made based on features, here decisions are made based on $\gamma_t^s(\mathbf{x})$, i.e., high-dimensional features are mapped into a one-dimensional space first and decisions are made in this transformed space. The threshold policy in this case becomes $\pi_t^s(\mathbf{x}) = \mathbf{1}(\gamma_t^s(\mathbf{x}) \geq \theta_t^s)$ with threshold $\theta_t^s \in [0, 1]$. Let $\gamma_t^{s-1}(\theta) \subset \mathbb{R}^d$ be defined as the preimage of θ under qualification profile γ_t^s , then all analysis in one-dimensional settings can be adjusted using $\gamma_t^{s-1}(\cdot)$. For example, Assumption 1 in high-dimensional case can be adjusted to the following: $\forall s \in \{a, b\}$, given any two thresholds $0 \leq \theta_j^s < \theta_k^s \leq 1$, we have $\gamma_t^{s-1}(K^s) \subset \gamma_t^{s-1}(J^s)$, where $J^s = \{\theta : \theta \in [\theta_j^s, 1]\}$ and $K^s = \{\theta : \theta \in [\theta_k^s, 1]\}$; in other words, if an individual can get accepted by a policy with the higher threshold, it must be accepted if a policy with a lower threshold was used. Note that this assumption is still mild and always hold if $G_y^s(\mathbf{x})$ belongs to exponential family.

Specifically, if $\forall s \in \{a, b\}, \forall y \in \{0, 1\}$, distribution of $X \mid Y = y, S = s$ belongs to exponential family and can be written as $G_y^s(\mathbf{x}) := B(\mathbf{x}) \exp(\langle \eta(\omega_y^s), \xi(\mathbf{x}) \rangle - A(\omega_y^s))$ for some functions $B(\cdot), \eta(\cdot), \xi(\cdot), A(\cdot)$, where $\langle \mathbf{x}, \mathbf{y} \rangle$ represents inner product of two vectors \mathbf{x}, \mathbf{y} and ω_y^s is the parameter. Then $\frac{G_0^s(\mathbf{x})}{G_1^s(\mathbf{x})} = \exp(-\langle \eta^s, \xi(\mathbf{x}) \rangle + A^s)$ where $\eta^s := \eta(\omega_1^s) - \eta(\omega_0^s)$ and $A^s := A(\omega_1^s) - A(\omega_0^s)$. Then

$$\gamma_t^{s-1}(J^s) = \{\mathbf{x} : \gamma_t^s(\mathbf{x}) \geq \theta_j^s\} = \{\mathbf{x} : \langle \eta^s, \xi(\mathbf{x}) \rangle \geq A^s + \log\left(\frac{\frac{1}{\alpha_t^s} - 1}{\frac{1}{\theta_j^s} - 1}\right)\}$$

If $\theta_j^s < \theta_k^s$, then $\log\left(\frac{\frac{1}{\alpha_t^s} - 1}{\frac{1}{\theta_j^s} - 1}\right) < \log\left(\frac{\frac{1}{\alpha_t^s} - 1}{\frac{1}{\theta_k^s} - 1}\right)$. We have $\gamma_t^{s-1}(K^s) \subset \gamma_t^{s-1}(J^s)$.

F Discussions

Transitions under Condition 1(C) or 1(D). This paper mainly focus on transitions satisfying Condition 1(A) and 1(B). As mentioned in Section 3.2, there are the other two combinations: (C)

$T_{01}^s \geq T_{00}^s$ and $T_{11}^s \leq T_{10}^s$; (D) $T_{01}^s \leq T_{00}^s$ and $T_{11}^s \geq T_{10}^s$, in which there is more uncertainty when conducting equilibrium analysis. The slight changes in the feature distributions or the values of transitions may change conclusions significantly.

Because the system has equilibrium if there is solution to balanced equations defined as Eqn. (5) in Appendix H, i.e., $\frac{1}{\alpha^s} - 1 = \frac{1-g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))}$, $\forall s \in \{a, b\}$. Since

$$\frac{1 - g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))} = \frac{1 - (T_{10}^s \mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b)) + T_{11}^s (1 - \mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b))))}{T_{00}^s \mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b)) + T_{01}^s (1 - \mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b)))}$$

Under optimal (fair) policies and Condition 1(A) or 1(B), $\frac{1-g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))}$ is guaranteed to be either decreasing or increasing in α^s . This monotonicity is critical to determine the properties (e.g., uniqueness, quantity, value, etc.) of the consequent equilibrium $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$ so that impacts of different fairness can be compared. In contrast, under Condition 1(C) or 1(D), $\frac{1-g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))}$ is no longer monotonic, and its intersection with function $\frac{1}{\alpha^s} - 1$, i.e., equilibrium, is thus hard to characterize. As a consequence, the impacts of different fairness constraints cannot be compared in general.

Comparison between sufficient conditions in Theorem 3 and Lipschitz condition. Let a pair of qualification rats of $\mathcal{G}_a, \mathcal{G}_b$ be noted as $\alpha := (\alpha^a, \alpha^b) \in [0, 1] \times [0, 1]$, and let mapping $\Phi : [0, 1] \times [0, 1] \rightarrow [0, 1] \times [0, 1]$ be defined such that dynamical system (4) can be written as $\alpha_{t+1} = \Phi(\alpha_t)$. Then this dynamical system has an equilibrium $\hat{\alpha}$ if $\Phi(\hat{\alpha}) = \hat{\alpha}$. According to Banach Fixed Point Theorem, such equilibrium exists and is unique if the mapping Φ satisfies L -Lipschitz condition with $L < 1$, i.e., Φ is a contraction mapping. Specifically, $d(\Phi(\alpha_0), \Phi(\alpha_1)) \leq Ld(\alpha_0, \alpha_1)$ for some distance function d and Lipschitz constant $L < 1$.

While Lipschitz condition also ensures the uniqueness of equilibrium, the sufficient conditions given in Theorem 3 are weaker. Use unconstrained optimal policies as an example, in this case dynamics of two groups can be decoupled because threshold $\theta^s(\alpha^a, \alpha^b)$ used in \mathcal{G}_s is independent of qualification of the other group α^{-s} . Therefore, sufficient condition $|\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^{-s}}| = 0 < 1$ under Condition 1(A) always holds. In contrast, for dynamics of \mathcal{G}_s after decoupling $\alpha_{t+1}^s = \Phi^s(\alpha_t^s) = g^{0s}(\theta^s(\alpha_t^s))(1 - \alpha_t^s) + g^{1s}(\theta^s(\alpha_t^s))\alpha_t^s$, Φ^s is not necessarily a contraction mapping.

Although sufficient conditions in Theorem 3 are weaker, they do not guarantee the stability of the equilibrium. In contrast, Lipschitz condition with $L < 1$ ensures the unique equilibrium is also stable, i.e., we have $(\alpha_t^a, \alpha_t^b) \rightarrow (\hat{\alpha}^a, \hat{\alpha}^b)$ given an arbitrary initial state (α_0^a, α_0^b) .

Impacts of estimation error in qualification profiles $\gamma_t^s(x)$. In the paper, the institute is assumed to have the perfect knowledge of $\gamma_t^s(x)$ and decisions are made based on it. In practice this can be obtained via learning/estimating $\alpha_t^s, G_y^s(x)$ from observable features and sensitive attributes. Many methods have been studied in the literature which can be adopted. Here we focus on examining the impacts of estimation error on the equilibrium of qualification rates. Specifically, we consider two types of error: (1) random error in estimating α_t^s for some t ; (2) under-estimating $\gamma_t^s(x)$ for disadvantaged group $\forall t$.

If the equilibria are stable, then they are robust to small perturbations that occur due to the first type of errors. Therefore, we focus on the second type of errors.

WLOG, let \mathcal{G}_b be labeled as disadvantaged group. We first consider the scenario where the \mathcal{G}_b 's qualification profile $\gamma_t^b(x)$ is under-estimated by the institute, potentially due to historical prejudice. We wonder how groups' qualification rates can be affected in the long-run when applying UN, DP, EqOpt policies. Let $\bar{\gamma}_t^s(x)$ be the estimated qualification profile of \mathcal{G}_s at t , and $(\bar{\alpha}_C^a, \bar{\alpha}_C^b)$ be the consequent equilibrium under $C = \text{UN, DP or EqOpt}$ policies. Let $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$ be equilibrium if true qualification profiles $\gamma_t^s(x)$ had been used when making decisions.

Proposition 4. Suppose $\gamma_t^b(x)$ is always under-estimated by institute, i.e., $\bar{\gamma}_t^b(x) \leq \gamma_t^b(x), \forall t, x$ but $\bar{\gamma}_t^a(x) = \gamma_t^a(x)$. We have,

- Under Condition 1(A), $\bar{\alpha}_{UN}^a = \hat{\alpha}_{UN}^a, \bar{\alpha}_{UN}^b \geq \hat{\alpha}_{UN}^b$; and $\bar{\alpha}_C^a \geq \hat{\alpha}_C^a, \bar{\alpha}_C^b \geq \hat{\alpha}_C^b$ for $C = \text{DP or EqOpt}$.
- Under Condition 1(B), $\bar{\alpha}_{UN}^a = \hat{\alpha}_{UN}^a, \bar{\alpha}_{UN}^b \leq \hat{\alpha}_{UN}^b$; and $\bar{\alpha}_C^a \leq \hat{\alpha}_C^a, \bar{\alpha}_C^b \leq \hat{\alpha}_C^b$ for $C = \text{DP or EqOpt}$.

Note that estimation error considered in this case is deterministic, i.e., given α_t^s , estimation $\bar{\gamma}_t^s(x)$ is deterministic and t -independent. While \mathcal{G}_a is not under-estimated, its equilibrium state is affected by \mathcal{G}_b when $\mathcal{C} = \text{DP}$ or EqOpt . Specifically, when $\gamma_t^b(x)$ is under-estimated, institute will under-select both groups to satisfy fairness.

G Impact of decisions on generation process

G.1 Results

In this section we consider the scenario where institute's decision not only affects the transition of the qualification state but also affects the generation process of individual feature X_t . We will only focus on analyzing its impact on equilibrium under unconstrained policy, the analysis only involves one group. Therefore, we remove the subscript $s \in \{a, b\}$ in this section. Let the generation probability distribution be $G_{yd}(x) := \mathbb{P}(X_t = x \mid Y_t = y, D_{t-1} = d)$ and corresponding CDF be $\mathbb{G}_{yd}(x)$.

Denote $\zeta_t^{yd} := \mathbb{P}(D_{t-1} = d, Y_t = y)$, then qualification profile $\gamma_t(x) = \frac{1}{1 + \frac{G_{00}(x)\zeta_t^{00} + G_{01}(x)\zeta_t^{01}}{G_{10}(x)\zeta_t^{10} + G_{11}(x)\zeta_t^{11}}}$ and

qualification rate $\alpha_t = \zeta_t^{11} + \zeta_t^{10}$. See Appendix G.2 for derivations and dynamics of α_t .

Assumption 3. $\frac{G_{00}(x)}{G_{10}(x)}, \frac{G_{00}(x)}{G_{11}(x)}, \frac{G_{01}(x)}{G_{10}(x)}, \frac{G_{01}(x)}{G_{11}(x)}$ are continuous and strictly decreasing in x .

Under Assumption 3, given $\{\zeta_t^{yd}\}_{y,d}$, $\gamma_t(x)$ is strictly increasing in x and unconstrained optimal policy is threshold policy. In the following we focus on a special case where decision only affect feature of (un)qualified people, i.e., $G_{01}(x) \neq G_{00}(x), G_{11}(x) = G_{10}(x) = G_1(x)$ or $G_{11}(x) \neq G_{10}(x), G_{01}(x) = G_{00}(x) = G_0(x)$. We wonder as compared to the case when $G_{y1}(x) = G_{y0}(x), \forall y = 0, 1$ (main body of the paper), how is the equilibrium under unconstrained optimal policies affected in these cases.

Assumption 4. $\mathbb{G}_{01}(x) \leq \mathbb{G}_{00}(x), \forall x$ and \exists a unique \hat{x} such that $G_{01}(\hat{x}) = G_{00}(\hat{x}), G_{01}(x) < G_{00}(x), \forall x < \hat{x}, G_{01}(x) > G_{00}(x), \forall x > \hat{x}$.

Assumption 5. $\mathbb{G}_{01}(x) \geq \mathbb{G}_{00}(x), \forall x$ and \exists a unique \hat{x} such that $G_{01}(\hat{x}) = G_{00}(\hat{x}), G_{01}(x) > G_{00}(x), \forall x < \hat{x}, G_{01}(x) < G_{00}(x), \forall x > \hat{x}$.

Theorem 7. Consider two scenarios with the same set of transitions $\{T_{yd}\}_{y,d}$ and the same $G_{11}(x) = G_{10}(x)$: scenario A has distributions $G_{00}^A(x) \neq G_{01}^A(x)$ and scenario B has $G_{00}^B(x) = G_{01}^B(x) = G_{01}^A(x)$. Let $\hat{\alpha}_B$ be the unique equilibrium of scenario B. Then under Assumption 3, scenario A has at least one equilibrium $\hat{\alpha}_A$. Moreover, if $\frac{u_+}{u_-} > \frac{G_{01}^A(\hat{x})}{G_{11}(\hat{x})} \max\{\frac{1-T_{10}}{T_{00}}, \frac{1-T_{11}}{T_{01}}\}$, we have:

- Under Assumption 4, we have $\hat{\alpha}_A > \hat{\alpha}_B$ under Condition 1(A); $\hat{\alpha}_A < \hat{\alpha}_B$ under Condition 1(B).
- Under Assumption 5, we have $\hat{\alpha}_A < \hat{\alpha}_B$ under Condition 1(A); $\hat{\alpha}_A > \hat{\alpha}_B$ under Condition 1(B).

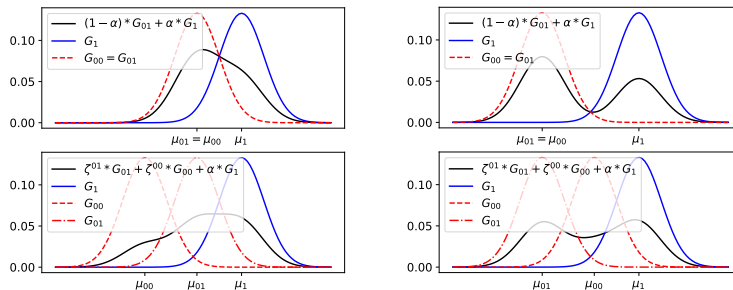


Figure 8: Illustration of feature distributions in two scenarios: the left plot shows the case when $\mathbb{G}_{01}^A(\theta) \leq \mathbb{G}_{00}^A(\theta), \forall \theta$, with $G_{00}^A(X) \sim \mathcal{N}(-8, 9), G_{01}^A(X) \sim \mathcal{N}(-1, 9)$; the right plot shows the case when $\mathbb{G}_{01}^A(\theta) \geq \mathbb{G}_{00}^A(\theta), \forall \theta$, with $G_{00}^A(X) \sim \mathcal{N}(-1, 9), G_{01}^A(X) \sim \mathcal{N}(-8, 9)$. Moreover, $G_1(X) \sim \mathcal{N}(5, 9), \zeta^{00} = 0.2, \alpha = 0.4$ in both cases. Within each plot, the upper and lower subplots show distributions in scenario B and A respectively.

Similar results can be concluded for case when $G_{11}(x) \neq G_{10}(x), G_{01}(x) = G_{00}(x) = G_0(x)$.

Assumption 6. $\mathbb{G}_{11}(x) \leq \mathbb{G}_{10}(x), \forall x$ and \exists a unique \hat{x} such that $G_{11}(\hat{x}) = G_{10}(\hat{x}), G_{11}(x) < G_{10}(x), \forall x < \hat{x}, G_{11}(x) > G_{10}(x), \forall x > \hat{x}$.

Assumption 7. $\mathbb{G}_{11}(x) \geq \mathbb{G}_{10}(x), \forall x$ and \exists a unique \hat{x} such that $G_{11}(\hat{x}) = G_{10}(\hat{x}), G_{11}(x) > G_{10}(x), \forall x < \hat{x}, G_{11}(x) < G_{10}(x), \forall x > \hat{x}$.

Theorem 8. Consider two scenarios with the same set of transitions $\{T_{yd}\}_{y,d}$ and the same $G_{01}(x) = G_{00}(x)$: scenario A has distributions $G_{10}^A(x) \neq G_{11}^A(x)$ and scenario B has $G_{10}^B(x) = G_{11}^B(x) = G_{10}^A(x)$. Let $\hat{\alpha}_B$ be the unique equilibrium of scenario B. Then under Assumption 3, scenario A has at least one equilibrium $\hat{\alpha}_A$. Moreover, if $\frac{u_+}{u_-} < \frac{G_{00}(\hat{x})}{G_{10}^A(\hat{x})} \min\{\frac{1-T_{10}}{T_{00}}, \frac{1-T_{11}}{T_{01}}\}$, we have

- Under Assumption 6, we have $\hat{\alpha}_A < \hat{\alpha}_B$ under Condition 1(A); $\hat{\alpha}_A > \hat{\alpha}_B$ under Condition 1(B).
- Under Assumption 7, we have $\hat{\alpha}_A > \hat{\alpha}_B$ under Condition 1(A); $\hat{\alpha}_A < \hat{\alpha}_B$ under Condition 1(B).

G.2 Derivations

Qualification profile.

$$\begin{aligned} \gamma_t(x) &= \mathbb{P}(Y_t = 1 | X_t = x) = \frac{1}{1 + \frac{\mathbb{P}(Y_t=0, X_t=x)}{\mathbb{P}(Y_t=1, X_t=x)}} \\ &= \frac{1}{1 + \frac{\sum_d \mathbb{P}(X_t=x | D_{t-1}=d, Y_t=0) \mathbb{P}(D_{t-1}=d, Y_t=0)}{\sum_d \mathbb{P}(X_t=x | D_{t-1}=d, Y_t=1) \mathbb{P}(D_{t-1}=d, Y_t=1)}} \\ &= \frac{1}{1 + \frac{G_{00}(x)\zeta_t^{00} + G_{01}(x)\zeta_t^{01}}{G_{10}(x)\zeta_t^{10} + G_{11}(x)\zeta_t^{11}}}. \end{aligned}$$

Dynamics of qualification rate

$$\begin{aligned} \zeta_t^{yd} &= \mathbb{P}(D_{t-1} = d, Y_t = y) \\ &= \sum_{d', y'} \mathbb{P}(Y_t = y, D_{t-1} = d | Y_{t-1} = y', D_{t-2} = d') \mathbb{P}(Y_{t-1} = y', D_{t-2} = d') \\ &= \sum_{d', y'} \mathbb{P}(Y_{t-1} = y', D_{t-2} = d') \mathbb{P}(Y_t = y | D_{t-1} = d, Y_{t-1} = y') \\ &\quad \mathbb{P}(D_{t-1} = d | Y_{t-1} = y', D_{t-2} = d') \\ &= \sum_{d', y'} \mathbb{P}(Y_{t-1} = y', D_{t-2} = d') \mathbb{P}(Y_t = y | D_{t-1} = d, Y_{t-1} = y') \\ &\quad \int \mathbb{P}(D_{t-1} = d | X_{t-1} = x') \mathbb{P}(X_{t-1} = x | Y_{t-1} = y', D_{t-2} = d') dx'. \end{aligned}$$

Simplify notation $\mathbb{E}_{X_{t-1}|Y_{t-1}=y', D_{t-2}=d'}[\cdot]$ as $\mathbb{E}^{y'd'}[\cdot]$, and $\pi_{t-1}(X_{t-1})$ as π . Evolution equation is given as:

$$\zeta_t = \Gamma_{t-1} \zeta_{t-1}.$$

where $\zeta_t = [\zeta_t^{11} \quad \zeta_t^{10} \quad \zeta_t^{01} \quad \zeta_t^{00}]^T$ and

$$\Gamma_{t-1} = \begin{bmatrix} T_{11}\mathbb{E}^{11}[\pi] & T_{11}\mathbb{E}^{10}[\pi] & T_{01}\mathbb{E}^{01}[\pi] & T_{01}\mathbb{E}^{00}[\pi] \\ T_{10}\mathbb{E}^{11}[1-\pi] & T_{10}\mathbb{E}^{10}[1-\pi] & T_{00}\mathbb{E}^{01}[1-\pi] & T_{00}\mathbb{E}^{00}[1-\pi] \\ (1-T_{11})\mathbb{E}^{11}[\pi] & (1-T_{11})\mathbb{E}^{10}[\pi] & (1-T_{01})\mathbb{E}^{01}[\pi] & (1-T_{01})\mathbb{E}^{00}[\pi] \\ ((1-T_{10})\mathbb{E}^{11}[1-\pi] & (1-T_{10})\mathbb{E}^{10}[1-\pi] & (1-T_{00})\mathbb{E}^{01}[1-\pi] & (1-T_{00})\mathbb{E}^{00}[1-\pi] \end{bmatrix}$$

Qualification rate satisfies $\alpha_t = \zeta_t^{11} + \zeta_t^{10}$.

H Proofs

We define balanced equations and functions for the rest proofs. The dynamics system (4) can reach equilibrium if $\alpha_t^s = \alpha_{t-1}^s$ holds. Therefore, the system has equilibrium if there exists solution to the *balanced equations* defined as (5).

$$\frac{1}{\alpha^a} - 1 = \frac{1 - g^{1a}(\theta^a(\alpha^a, \alpha^b))}{g^{0a}(\theta^a(\alpha^a, \alpha^b))}; \quad \frac{1}{\alpha^b} - 1 = \frac{1 - g^{1b}(\theta^b(\alpha^a, \alpha^b))}{g^{0b}(\theta^b(\alpha^a, \alpha^b))}. \quad (5)$$

By removing subscript t and re-writing threshold θ^s as a function of α^a, α^b , the notation g_t^{ys} in (4) is denoted as $g^{ys}(\theta^s(\alpha^a, \alpha^b)) = T_{y0}^s \mathbb{G}_y^s(\theta^s(\alpha^a, \alpha^b)) + T_{y1}^s(1 - \mathbb{G}_y^s(\theta^s(\alpha^a, \alpha^b)))$ in (5).

$\forall s \in \{a, b\}$, let $-s := \{a, b\} \setminus s$. $\forall \alpha^{-s} \in [0, 1]$, define *balanced set* w.r.t. dynamics as $\Psi^s(\alpha^{-s}) := \{\bar{\alpha}^s : \frac{1}{\bar{\alpha}^s} - 1 = \frac{1 - g^{1s}(\theta^a(\bar{\alpha}^s, \alpha^{-s}))}{g^{0s}(\theta^s(\bar{\alpha}^s, \alpha^{-s}))}\}$. If the set size $|\Psi^s(\alpha^{-s})| = 1$ holds $\forall \alpha^{-s} \in [0, 1]$, we define *balanced functions* w.r.t. dynamics as $\psi^s : [0, 1] \rightarrow [0, 1]$ with $\psi^s(\alpha^{-s}) \in \Psi^s(\alpha^{-s}), \forall \alpha^{-s} \in [0, 1]$.

The proof of Theorem 1.

Proof. In the following proof, we focus on optimal policy at t and omit the subscript t .

First consider unconstrained optimal policy, noted as π_{UN}^s , we have,

$$\pi_{\text{UN}}^s = \arg \max_{\pi^s} \mathbb{E}_{X|S=s}[\pi^s(X)(\gamma^s(X)(u_+ + u_-) - u_-)]$$

Therefore, the optimal policy satisfies $\pi_{\text{UN}}^s(x) = \mathbf{1}(\gamma^s(x) \geq \frac{u_-}{u_+ + u_-})$. Since $\gamma^s(x)$ is monotonically increasing in x under Assumption 1, $\pi_{\text{UN}}^s(x) = \mathbf{1}(x \geq (\gamma^s)^{-1}(\frac{u_-}{u_+ + u_-}))$ is threshold policy where $(\gamma^s)^{-1}(\cdot)$ denotes the inverse function of $\gamma(\cdot)$.

Now consider optimal fair policy under some fairness constraint \mathcal{C} satisfying Assumption 2. Consider any pair of policies (π^a, π^b) that satisfies fairness constraint \mathcal{C} , and define fairness constant $c = \mathbb{E}_{X \sim \mathcal{P}_c^a}[\pi^a(X)] = \mathbb{E}_{X \sim \mathcal{P}_c^b}[\pi^b(X)] \in [0, 1]$. To show the optimal fair policy is threshold policy, we will show that there always exists a pair of threshold policies (π_d^a, π_d^b) such that $\mathbb{E}_{X \sim \mathcal{P}_c^a}[\pi_d^a(X)] = \mathbb{E}_{X \sim \mathcal{P}_c^b}[\pi_d^b(X)] = c$, i.e., the fairness constant is the same as (π^a, π^b) , and the utility of (π_d^a, π_d^b) is no less than the utility attained under (π^a, π^b) .

$\forall s \in \{a, b\}$, let threshold policy π_d^s be defined such that $\pi_d^s(x) = \mathbf{1}(x \geq \theta_d^s)$ and $\mathbb{E}_{X \sim \mathcal{P}_c^s}[\pi_d^s(X)] = c$ are satisfied. Such policy must exist and the threshold is given by $\theta_d^s = (\mathbb{P}_c^s)^{-1}(1 - c)$, where $\mathbb{P}_c^s(\theta^s) = \int_{-\infty}^{\theta^s} \mathcal{P}_c^s(x) dx$ is CDF of \mathcal{P}_c^s and $(\mathbb{P}_c^s)^{-1}(\cdot)$ is the inverse of it.

Let $R_{\pi_d^s}(D, Y), R_{\pi^s}(D, Y)$ denote the utility attained under policies π_d^s, π^s respectively. Next we will show that $\forall s \in \{a, b\}, \mathbb{E}[R_{\pi_d^s}(D, Y) | S = s] \geq \mathbb{E}[R_{\pi^s}(D, Y) | S = s]$ holds, i.e.,

$$\mathbb{E}_{X|S=s}[\pi_d^s(X)(\gamma^s(X)(u_+ + u_-) - u_-)] \geq \mathbb{E}_{X|S=s}[\pi^s(X)(\gamma^s(X)(u_+ + u_-) - u_-)]$$

Since $\pi_d^s(x) = \mathbf{1}(x \geq \theta_d^s)$, we have the followings,

$$\begin{aligned} \mathbb{E}_{X|S=s}[\pi_d^s(X)(\gamma^s(X)(u_+ + u_-) - u_-)] &= \int_{\theta_d^s}^{\infty} (\gamma^s(x)(u_+ + u_-) - u_-) \mathbb{P}(X = x | S = s) dx \\ \mathbb{E}_{X|S=s}[\pi^s(X)(\gamma^s(X)(u_+ + u_-) - u_-)] &= \int_{\theta_d^s}^{\infty} (\gamma^s(x)(u_+ + u_-) - u_-) \mathbb{P}(X = x | S = s) dx \\ &\quad + \int_{-\infty}^{\theta_d^s} \pi^s(x)(\gamma^s(x)(u_+ + u_-) - u_-) \mathbb{P}(X = x | S = s) dx \\ &\quad - \int_{\theta_d^s}^{\infty} (1 - \pi^s(x))(\gamma^s(x)(u_+ + u_-) - u_-) \mathbb{P}(X = x | S = s) dx \end{aligned}$$

Since $\mathbb{E}_{X \sim \mathcal{P}_c^s}[\pi^s(X)] = c = \mathbb{E}_{X \sim \mathcal{P}_c^s}[\pi_d^s(X)]$, we have

$$\int_{\theta_d^s}^{\infty} (1 - \pi^s(x)) \mathcal{P}_c^s(x) dx = \int_{-\infty}^{\theta_d^s} \pi^s(x) \mathcal{P}_c^s(x) dx \quad (6)$$

Under Assumption 2, $\frac{G_1^s(x)}{\mathcal{P}_c^s(x)}$ is non-decreasing and $\frac{G_0^s(x)}{\mathcal{P}_c^s(x)}$ is non-increasing. Therefore,

$$(\gamma^s(x)(u_+ + u_-) - u_-) \frac{\mathbb{P}(X = x | S = s)}{\mathcal{P}_c^s(x)} = \alpha^s \frac{G_1^s(x)}{\mathcal{P}_c^s(x)} u_+ - (1 - \alpha^s) \frac{G_0^s(x)}{\mathcal{P}_c^s(x)} u_-,$$

is non-decreasing in x . Combine with Eqn. (6), we have the followings,

$$\begin{aligned}
& \int_{-\infty}^{\theta_d^s} \pi^s(x)(\gamma^s(x)(u_+ + u_-) - u_-)\mathbb{P}(X = x | S = s)dx \\
&= \int_{-\infty}^{\theta_d^s} \pi^s(x)(\gamma^s(x)(u_+ + u_-) - u_-)\frac{\mathbb{P}(X = x | S = s)}{\mathcal{P}_{\mathcal{C}}^s(x)}\mathcal{P}_{\mathcal{C}}^s(x)dx \\
&\leq \int_{-\infty}^{\theta_d^s} \pi^s(x)(\gamma^s(\theta_d^s)(u_+ + u_-) - u_-)\frac{\mathbb{P}(X = \theta_d^s | S = s)}{\mathcal{P}_{\mathcal{C}}^s(\theta_d^s)}\mathcal{P}_{\mathcal{C}}^s(x)dx \\
&= \int_{\theta_d^s}^{\infty} (1 - \pi^s(x))(\gamma^s(\theta_d^s)(u_+ + u_-) - u_-)\frac{\mathbb{P}(X = \theta_d^s | S = s)}{\mathcal{P}_{\mathcal{C}}^s(\theta_d^s)}\mathcal{P}_{\mathcal{C}}^s(x)dx \\
&\leq \int_{\theta_d^s}^{\infty} (1 - \pi^s(x))(\gamma^s(x)(u_+ + u_-) - u_-)\frac{\mathbb{P}(X = x | S = s)}{\mathcal{P}_{\mathcal{C}}^s(x)}\mathcal{P}_{\mathcal{C}}^s(x)dx \\
&= \int_{\theta_d^s}^{\infty} (1 - \pi^s(x))(\gamma^s(x)(u_+ + u_-) - u_-)\mathbb{P}(X = x | S = s)dx.
\end{aligned}$$

Therefore, the following holds $\forall s \in \{a, b\}$,

$$\mathbb{E}_{X|S=s}[\pi_d^s(X)(\gamma^s(X)(u_+ + u_-) - u_-)] \geq \mathbb{E}_{X|S=s}[\pi^s(X)(\gamma^s(X)(u_+ + u_-) - u_-)].$$

It shows that the utility attained under threshold policy (π_d^a, π_d^b) is no less than the utility of (π^a, π^b) , which concludes that the optimal fair policy $(\pi_{\mathcal{C}}^a, \pi_{\mathcal{C}}^b)$ must be threshold policies.

Lemma 2 below further shows that the optimal threshold policy $\theta^s(\alpha^a, \alpha^b)$ is continuous and non-increasing in α^a and α^b .

Lemma 2. *Let $(\theta^a(\alpha^a, \alpha^b), \theta^b(\alpha^a, \alpha^b))$ be a pair of solutions to Eqn. (3) under α^a, α^b . $\forall s \in \{a, b\}$, if $\frac{G_1^s(x)}{\mathcal{P}_{\mathcal{C}}^s(x)}$ and $\frac{G_0^s(x)}{\mathcal{P}_{\mathcal{C}}^s(x)}$ are continuous everywhere in x , then $\theta^s(\alpha^a, \alpha^b)$ is continuous in both α^a and α^b . Moreover, under Assumption 2, $\theta^s(\alpha^a, \alpha^b)$ is non-increasing in α^a and α^b .*

Proof. To prove that a sufficient condition under which $\theta^s(\alpha^a, \alpha^b)$ is continuous in $\alpha^a, \alpha^b \in [0, 1]$ is that $\frac{G_1^s(x)}{\mathcal{P}_{\mathcal{C}}^s(x)}$ and $\frac{G_0^s(x)}{\mathcal{P}_{\mathcal{C}}^s(x)}$ are continuous everywhere in x , we define a function $f^s(\theta^s, \alpha^a, \alpha^b)$:

$$\begin{aligned}
f^s(\theta^s, \alpha^a, \alpha^b) &= (\gamma^s(\theta^s) - \frac{u_-}{u_+ + u_-})\frac{\mathbb{P}(X = \theta^s | S = s)}{\mathcal{P}_{\mathcal{C}}^s(\theta^s)} \\
&= [\alpha^s u_+ G_1^s(\theta^s) + \alpha^s u_- G_0^s(\theta^s) - u_- G_0^s(\theta^s)]\frac{1}{\mathcal{P}_{\mathcal{C}}^s(\theta^s)} \\
&= [\alpha^s \frac{G_1^s(\theta^s)}{\mathcal{P}_{\mathcal{C}}^s(\theta^s)} u_+ + (\alpha^s - 1)\frac{G_0^s(\theta^s)}{\mathcal{P}_{\mathcal{C}}^s(\theta^s)} u_-].
\end{aligned}$$

According to Equation (3), we have $p_a f^a(\theta^a, \alpha^a, \alpha^b) + p_b f^b(\theta^b, \alpha^a, \alpha^b) = 0$.

Given any α^a and α^b , and any constant k , let $\tilde{\theta}_i^s$ be one solution to $f^s(\theta^s, \alpha^a, \alpha^b) = k$, where $i = 1, \dots, N$ and N is the number of solutions. Firstly, we show that $\tilde{\theta}_i^s(\alpha^a, \alpha^b)$ is continuous in α^a and α^b , for any $i \in \{1, \dots, N\}$. Because $\frac{G_1^s(x)}{\mathcal{P}_{\mathcal{C}}^s(x)}$ and $\frac{G_0^s(x)}{\mathcal{P}_{\mathcal{C}}^s(x)}$ are continuous, $f^s(\theta^s, \alpha^a, \alpha^b)$ is continuous in α^a, α^b , and θ^s . Therefore, $\forall \epsilon > 0, \exists \delta > 0$ such that for all $|\alpha^{a'} - \alpha^a| < \delta$ and $|\alpha^{b'} - \alpha^b| < \delta \implies |\tilde{\theta}_i^{s'} - \tilde{\theta}_i^s| < \epsilon$. Thus, $\tilde{\theta}_i^s(\alpha^a, \alpha^b)$ is continuous in α^a and $\alpha^b, \forall i \in \{1, \dots, N\}$.

Next, we show that given α^a and α^b , the solutions to $p_a f^a(\theta^a, \alpha^a, \alpha^b) + p_b f^b(\theta^b, \alpha^a, \alpha^b) = 0$ under fairness constraint \mathcal{C} are continuous in α^a and $\alpha^b \in [0, 1]$. Under fairness constraints in Equation (1), $\theta^a = \phi_{\mathcal{C}}(\theta^b)$ holds for some continuous function $\phi_{\mathcal{C}}(\cdot)$. Consequently, we have $p_a f^a(\phi_{\mathcal{C}}(\theta^b), \alpha^a, \alpha^b) + p_b f^b(\theta^b, \alpha^a, \alpha^b) = 0$. Because $f^s(\cdot, \cdot, \cdot)$ and $\phi_{\mathcal{C}}(\cdot)$ are continuous functions, with the same reasoning, given α^a and α^b , the solutions to $p_a f^a(\phi_{\mathcal{C}}(\theta^b), \alpha^a, \alpha^b) + p_b f^b(\theta^b, \alpha^a, \alpha^b) = 0$ are continuous in α^a and α^b . In other words, $\theta_i^s(\alpha^a, \alpha^b)$ is continuous.

Under Assumption 2, $f^s(\theta^s, \alpha^a, \alpha^b)$ and $\theta^s(\alpha^a, \alpha^b)$ are continuous. We then prove that if $\frac{G_1^s(x)}{\mathcal{P}_C^s(x)}$ is non-decreasing and $\frac{G_0^s(x)}{\mathcal{P}_C^s(x)}$ is non-increasing in x , then $\theta^s(\alpha^a, \alpha^b)$ is non-increasing in α^a and α^b .

Let $(\phi_C(\theta^b), \theta^b)$ be a pair that satisfies fairness constraint, where $\phi_C(\cdot)$ is some continuous and strictly increasing function, then the optimal one is the pair that satisfies Equation (3) as follows:

$$\begin{aligned} & p_a \left(\gamma^a(\phi_C(\theta^b)) - \frac{u_-}{u_+ + u_-} \right) \frac{\mathbb{P}(X = \phi_C(\theta^b) | S = a)}{\mathcal{P}_C^a(\phi_C(\theta^b))} + p_b \left(\gamma^b(\theta^b) - \frac{u_-}{u_+ + u_-} \right) \frac{\mathbb{P}(X = \theta^b | S = b)}{\mathcal{P}_C^b(\theta^b)} \\ &= p_a \left[\alpha^a \frac{G_1^a(\phi_C(\theta^b))}{\mathcal{P}_C^a(\phi_C(\theta^b))} u_+ + (\alpha^a - 1) \frac{G_0^a(\phi_C(\theta^b))}{\mathcal{P}_C^a(\phi_C(\theta^b))} u_- \right] + p_b \left[\alpha^b \frac{G_1^b(\theta^b)}{\mathcal{P}_C^b(\theta^b)} u_+ + (\alpha^b - 1) \frac{G_0^b(\theta^b)}{\mathcal{P}_C^b(\theta^b)} u_- \right] = 0. \end{aligned}$$

Note that $\forall s \in \{a, b\}$, LHS of above equation is strictly increasing in α^s because the coefficient of α^s is positive. Because $\frac{G_1^s(x)}{\mathcal{P}_C^s(x)}$ is non-decreasing and $\frac{G_0^s(x)}{\mathcal{P}_C^s(x)}$ is non-increasing in x , $\frac{G_1^s(x)}{\mathcal{P}_C^s(x)} - \frac{G_0^s(x)}{\mathcal{P}_C^s(x)}$ is non-decreasing in x . As α^s increases, both $\frac{G_1^a(\phi_C(\theta^b))}{\mathcal{P}_C^a(\phi_C(\theta^b))} - \frac{G_0^a(\phi_C(\theta^b))}{\mathcal{P}_C^a(\phi_C(\theta^b))}$ and $\frac{G_1^b(\theta^b)}{\mathcal{P}_C^b(\theta^b)} - \frac{G_0^b(\theta^b)}{\mathcal{P}_C^b(\theta^b)}$ must not increase so that the optimal fair equation can be maintained. It requires that both θ^b and $\theta^a = \phi_C(\theta^b)$ must not increase. In other words, $\forall s \in \{a, b\}$, $\theta^s(\alpha^a, \alpha^b)$ must be non-increasing in α^a and α^b . \square

The proof of Lemma 1.

Proof. In the following proof, we focus on optimal policy at t and omit the subscript t .

First consider unconstrained optimal policy. Under threshold policy,

$$\begin{aligned} \theta_{\text{UN}}^{s*} &= \arg \max_{\theta^s} \mathbb{E}_{X|S=s} [\pi^s(X) (\gamma^s(X) (u_+ + u_-) - u_-)] \\ &= \arg \max_{\theta^s} \int_{\theta^s}^{\infty} (\gamma^s(x) (u_+ + u_-) - u_-) \mathbb{P}(X = x | S = s) dx \end{aligned}$$

Since $\gamma^s(x)$ is monotonically increasing in x under Assumption 1, θ_{UN}^{s*} satisfies $\gamma^s(\theta_{\text{UN}}^{s*}) = \frac{u_-}{u_+ + u_-}$.

Now consider optimal policy under fairness constraint, to satisfy constraint \mathcal{C} , $\int_{\theta^a}^{\infty} \mathcal{P}_C^a(x) dx = \int_{\theta^b}^{\infty} \mathcal{P}_C^b(x) dx$ should hold. Denote CDF $\mathbb{P}_C^s(\theta^s) = \int_{-\infty}^{\theta^s} \mathcal{P}_C^s(x) dx$, then for any pair (θ^a, θ^b) that is fair, we have $\theta^a = (\mathbb{P}_C^a)^{-1} \mathbb{P}_C^b(\theta^b) = \phi_C(\theta^b)$ hold for some strictly increasing function $\phi_C(\cdot)$. Denote $u = \mathbb{P}_C^b(\theta^b)$ and $\theta^a = (\mathbb{P}_C^a)^{-1}(u)$, the following holds,

$$\frac{d\phi_C(\theta^b)}{d\theta^b} = \frac{d(\mathbb{P}_C^a)^{-1}(\mathbb{P}_C^b(\theta^b))}{d\theta^b} = \frac{d(\mathbb{P}_C^a)^{-1}(u)}{du} \frac{du}{d\theta^b} = \frac{1}{(\mathbb{P}_C^a)'((\mathbb{P}_C^a)^{-1}(u))} \frac{du}{d\theta^b} = \frac{(\mathbb{P}_C^b)'(\theta^b)}{(\mathbb{P}_C^a)'(\theta^a)} = \frac{\mathcal{P}_C^b(\theta^b)}{\mathcal{P}_C^a(\theta^a)}.$$

Denote $f^s(x) := (\gamma^s(x) (u_+ + u_-) - u_-) \mathbb{P}(X = x | S = s)$, then we have

$$\theta_C^{b*} = \arg \max_{\theta^b} \mathcal{U}(D, Y) = \arg \max_{\theta^b} \left(p_a \int_{\phi_C(\theta^b)}^{\infty} f^a(x) dx + p_b \int_{\theta^b}^{\infty} f^b(x) dx \right).$$

Let $F(\theta^b) := p_a \int_{\phi_C(\theta^b)}^{\infty} f^a(x) dx + p_b \int_{\theta^b}^{\infty} f^b(x) dx$. Because $\gamma^s(x)$ is monotonically increasing in x under Assumption 1, the optimal θ_C^{b*} satisfies

$$\begin{aligned} \left. \frac{dF(\theta^b)}{d\theta^b} \right|_{\theta^b = \theta_C^{b*}} &= -p_a f^a(\phi_C(\theta^b)) \frac{d\phi_C(\theta^b)}{d\theta^b} - p_b f^b(\theta^b) \Big|_{\theta^b = \theta_C^{b*}} \\ &= -p_a (\gamma^a(\phi_C(\theta_C^{b*})) (u_+ + u_-) - u_-) \mathbb{P}(X = \phi_C(\theta_C^{b*}) | S = a) \frac{\mathcal{P}_C^b(\theta_C^{b*})}{\mathcal{P}_C^a(\phi_C(\theta_C^{b*}))} \\ &\quad - p_b (\gamma^b(\theta_C^{b*}) (u_+ + u_-) - u_-) \mathbb{P}(X = \theta_C^{b*} | S = b) \\ &= 0. \end{aligned}$$

Therefore,

$$p_a (\gamma^a(\theta_C^{a*}) (u_+ + u_-) - u_-) \frac{\mathbb{P}(X = \theta_C^{a*} | S = a)}{\mathcal{P}_C(\theta_C^{a*})} + p_b (\gamma^b(\theta_C^{b*}) (u_+ + u_-) - u_-) \frac{\mathbb{P}(X = \theta_C^{b*} | S = b)}{\mathcal{P}_C(\theta_C^{b*})} = 0. \quad \square$$

The proof of Theorem 2.

Proof. $\forall s \in \{a, b\}$, define function $l^s(\alpha^s) := \frac{1}{\alpha^s} - 1$ and $h^s(\theta^s(\alpha^a, \alpha^b)) := \frac{1 - g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))}$,

$$h^s(\theta^s(\alpha^a, \alpha^b)) = \frac{1 - (T_{10}^s \mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b)) + T_{11}^s(1 - \mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b))))}{T_{00}^s \mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b)) + T_{01}^s(1 - \mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b)))}.$$

Firstly, we prove that given a fixed $\alpha^{-s} \in [0, 1]$ there must exist at least one $\bar{\alpha}^s \in (0, 1)$ such that $h^s(\theta^s(\alpha^{-s}, \bar{\alpha}^s)) = l^s(\bar{\alpha}^s)$, $s \in \{a, b\}$, $-s = \{a, b\} \setminus s$.

Since $\mathbb{G}_y^s(x)$ is continuous in x , and $\theta^s(\alpha^a, \alpha^b)$ is continuous in α^a and α^b , $\mathbb{G}_y^s(\theta^s(\alpha^a, \alpha^b))$ is continuous in α^a and α^b . Therefore, $h^s(\theta^s(\alpha^a, \alpha^b))$ is continuous in α^a and α^b .

Moreover, $g^{1s}(\theta^s(\alpha^a, \alpha^b))$ is the convex combination of T_{11}^s and T_{10}^s , and $g^{0s}(\theta^s(\alpha^a, \alpha^b))$ is the convex combination of T_{01}^s and T_{00}^s , the following holds $\forall \alpha^a \in [0, 1], \alpha^b \in [0, 1]$,

$$\begin{aligned} \min\{T_{10}^s, T_{11}^s\} &\leq g^{1s}(\theta^s(\alpha^a, \alpha^b)) \leq \max\{T_{10}^s, T_{11}^s\}; \\ \min\{T_{00}^s, T_{01}^s\} &\leq g^{0s}(\theta^s(\alpha^a, \alpha^b)) \leq \max\{T_{00}^s, T_{01}^s\}, \end{aligned}$$

which implies $0 < \frac{1 - \max\{T_{10}^s, T_{11}^s\}}{\max\{T_{00}^s, T_{01}^s\}} \leq h^s(\theta^s(\alpha^a, \alpha^b)) \leq \frac{1 - \min\{T_{10}^s, T_{11}^s\}}{\min\{T_{00}^s, T_{01}^s\}} < +\infty$.

Furthermore, $l^s(\alpha^s) := \frac{1}{\alpha^s} - 1$ is continuous and strictly decreasing in α^s , and

$$\lim_{\alpha^s \rightarrow 0} l^s(\alpha^s) = +\infty; \quad \lim_{\alpha^s \rightarrow 1} l^s(\alpha^s) = 0,$$

Given a fixed $\alpha^a \in [0, 1]$, because $h^b(\theta^b(\alpha^a, \alpha^b))$ is continuous over $\alpha^b \in [0, 1]$ and with value varying between $\frac{1 - \max\{T_{10}^b, T_{11}^b\}}{\max\{T_{00}^b, T_{01}^b\}}$ and $\frac{1 - \min\{T_{10}^b, T_{11}^b\}}{\min\{T_{00}^b, T_{01}^b\}}$, and $l^b(\alpha^b)$ is continuous with value varying from $+\infty$ to 0, there must exist at least one $\bar{\alpha}^b \in (0, 1)$ such that $h^b(\theta^b(\alpha^a, \bar{\alpha}^b)) = l^b(\bar{\alpha}^b)$. Similarly, given a fixed $\alpha^b \in [0, 1]$, there must exist at least one $\bar{\alpha}^a \in (0, 1)$ such that $h^a(\theta^a(\bar{\alpha}^a, \alpha^b)) = l^a(\bar{\alpha}^a)$.

Secondly, we prove that all the solutions $(\bar{\alpha}^a, \alpha^b)$ and $(\alpha^a, \bar{\alpha}^b)$ are on continuous curves in the 2D plane $\{(\alpha^a, \alpha^b) : \alpha^a \in [0, 1], \alpha^b \in [0, 1]\}$.

According to the continuity of $l^s(\cdot)$ and $h^s(\cdot)$, we have $\forall \alpha^a \in [0, 1], \lim_{\alpha^{a'} \rightarrow \alpha^a} l^a(\alpha^{a'}) = l^a(\alpha^a)$; furthermore, $\forall \alpha^a \in [0, 1]$ and $\forall \theta_i^a \in \{\theta^a : l^a(\alpha^a) = h^a(\theta^a)\}$, $\lim_{\theta_i^{a'} \rightarrow \theta_i^a} h^a(\theta_i^{a'}) = h^a(\theta_i^a)$. Thus, $\forall \epsilon > 0, \exists \delta > 0$, such that $\forall \alpha^a \in [0, 1], |\alpha^{a'} - \alpha^a| < \delta \implies |\theta_i^{a'} - \theta_i^a| < \epsilon$. Consequently, $\forall \epsilon > 0, \exists \delta' > 0$ and $\exists \delta > 0$, such that $\forall \alpha^a \in [0, 1], |\alpha^{a'} - \alpha^a| < \delta \implies |\theta_i^{a'} - \theta_i^a| < \delta' \implies |\alpha_i^{b'} - \alpha_i^b| < \epsilon$, the last statement is because of the continuity of $\theta^a(\alpha^a, \alpha^b)$; in other words, $\forall \alpha^a \in [0, 1], \lim_{\alpha^{a'} \rightarrow \alpha^a} \alpha_i^{b'} = \alpha_i^b$, where $i = 1, \dots, N$. Therefore, $(\bar{\alpha}^a, \alpha^b)$ is on a set of continuous curves with α^b varying from 0 to 1. Similarly, one can prove that $(\alpha^a, \bar{\alpha}^b)$ is also on a set of continuous curves with α^a varying from 0 to 1.

Finally, we show the existence of equilibrium $(\hat{\alpha}^a, \hat{\alpha}^b)$.

Consider a 2D plane $\{(\alpha^a, \alpha^b) : \alpha^a \in [0, 1], \alpha^b \in [0, 1]\}$, and $\mathcal{C}_1 = \{(\bar{\alpha}^a, \alpha^b)\}$ and $\mathcal{C}_2 = \{(\alpha^a, \bar{\alpha}^b)\}$ that are two sets of continuous curves in the plane defined earlier. It is straightforward to see that there is at least one curve among \mathcal{C}_1 whose α^b varies from 0 to 1 and at least one curve among \mathcal{C}_2 whose α^a varies from 0 to 1. These two continuous curves must have at least one intersection. Moreover, this intersection $(\hat{\alpha}^a, \hat{\alpha}^b)$ satisfies $h^b(\theta^b(\hat{\alpha}^a, \hat{\alpha}^b)) = l^b(\hat{\alpha}^b)$ and $h^a(\theta^a(\hat{\alpha}^a, \hat{\alpha}^b)) = l^a(\hat{\alpha}^a)$, is an equilibrium of system.

Moreover, we also realized that the proof can also be done by using Brouwer's Fixed Point Theorem in topology. \square

Proof of Theorem 3

Proof. Following the proof of Theorem 2,

$$h^s(\theta^s(\alpha^a, \alpha^b)) = \frac{1 - g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))} = \frac{1 - (T_{10}^s \mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b)) + T_{11}^s(1 - \mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b))))}{T_{00}^s \mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b)) + T_{01}^s(1 - \mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b)))}.$$

Note that $\forall y \in \{0, 1\}$, $T_{y0}^s \mathbb{G}_y^s(\theta^s(\alpha^a, \alpha^b)) + T_{y1}^s(1 - \mathbb{G}_y^s(\theta^s(\alpha^a, \alpha^b)))$ is the convex combination of T_{y0}^s and T_{y1}^s with CDF $\mathbb{G}_y^s(\theta^s(\alpha^a, \alpha^b))$ as the weight. Because $\mathbb{G}_y^s(\theta^s(\alpha^a, \alpha^b))$ is continuous and non-decreasing in $\theta^s(\alpha^a, \alpha^b)$, under Condition 1(A), $h^s(\theta^s(\alpha^a, \alpha^b))$ is non-decreasing in $\theta^s(\alpha^a, \alpha^b)$; while under Condition 1(B), $h^s(\theta^s(\alpha^a, \alpha^b))$ is non-increasing in $\theta^s(\alpha^a, \alpha^b)$.

Under unconstrained optimal policy or optimal fair policy with constraint \mathcal{C} satisfying Assumption 1 and 2, $\theta^s(\alpha^a, \alpha^b)$ is non-increasing in α^a, α^b . Therefore, under Condition 1(A), $h^s(\theta^s(\alpha^a, \alpha^b))$ is non-decreasing in α^a, α^b , while under Condition 1(B), $h^s(\theta^s(\alpha^a, \alpha^b))$ is non-increasing in α^a, α^b . Moreover,

$$\text{Under Condition 1(A):} \quad 0 < \frac{1 - T_{10}^s}{T_{00}^s} \leq h^s(\theta^s(\alpha^a, \alpha^b)) \leq \frac{1 - T_{11}^s}{T_{01}^s} < +\infty$$

$$\text{Under Condition 1(B):} \quad 0 < \frac{1 - T_{11}^s}{T_{01}^s} \leq h^s(\theta^s(\alpha^a, \alpha^b)) \leq \frac{1 - T_{10}^s}{T_{00}^s} < +\infty$$

First consider the case when Condition 1(A) is satisfied.

Because function $l^s(\alpha^s) := \frac{1}{\alpha^s} - 1$ is continuous and strictly decreasing from $+\infty$ to 0 over $\alpha^s \in [0, 1]$, $\forall s \in \{a, b\}$. Thus, given any fixed $\alpha^b \in [0, 1]$, strictly decreasing function $l^a(\alpha^a)$ and non-decreasing function $h^a(\theta^a(\alpha^a, \alpha^b))$ has exactly one intersection, i.e., \exists only one $\bar{\alpha}^a$ such that $h^a(\theta^a(\bar{\alpha}^a, \alpha^b)) = l^a(\bar{\alpha}^a)$. $\forall \alpha^b$, the set $\Psi^a(\alpha^b) = \{\bar{\alpha}^a : h^a(\theta^a(\bar{\alpha}^a, \alpha^b)) = l^a(\bar{\alpha}^a)\}$ has only one element, and they constitute continuous function $\bar{\alpha}^a = \psi^a(\alpha^b)$ (balanced function). Similarly, $\forall \alpha^a$, set $\Psi^b(\alpha^a) = \{\bar{\alpha}^b : h^b(\theta^b(\alpha^a, \bar{\alpha}^b)) = l^b(\bar{\alpha}^b)\}$ also has only one element, which forms continuous function $\bar{\alpha}^b = \psi^b(\alpha^a)$.

Because given any α^a , $h^a(\theta^a(\alpha^a, \alpha^b))$ is non-decreasing in α^b , as α^b increases, the intersection with $l^a(\bar{\alpha}^a)$ is non-increasing. Therefore, $\psi^a(\alpha^b)$ is non-increasing in α^b . Similarly, $\psi^b(\alpha^a)$ is also non-increasing in α^a .

On the 2D plane $\{(\alpha^a, \alpha^b) : \alpha^a \in [0, 1], \alpha^b \in [0, 1]\}$, two curves $\mathcal{C}_1 = \{(\alpha^a, \alpha^b) : \alpha^a = \psi^a(\alpha^b), \alpha^b \in [0, 1]\}$ and $\mathcal{C}_2 = \{(\alpha^a, \alpha^b) : \alpha^b = \psi^b(\alpha^a), \alpha^a \in [0, 1]\}$ are both continuous and non-increasing. One sufficient condition to guarantee \mathcal{C}_1 and \mathcal{C}_2 have exact one intersection, is that $|\frac{d\psi^a(\alpha^b)}{d\alpha^b}| < 1, \forall \alpha^b \in [0, 1]$ and $|\frac{d\psi^b(\alpha^a)}{d\alpha^a}| < 1, \forall \alpha^a \in [0, 1]$. In the followings, we show these sufficient conditions will hold if $|\frac{\partial h^a(\theta^a(\alpha^a, \alpha^b))}{\partial \alpha^b}| < 1$ and $|\frac{\partial h^b(\theta^b(\alpha^a, \alpha^b))}{\partial \alpha^a}| < 1, \forall \alpha^a, \alpha^b$.

Denote $u := h^a(\theta^a(\psi^a(\alpha^b), \alpha^b))$, because $l^a(\psi^a(\alpha^b)) = h^a(\theta^a(\psi^a(\alpha^b), \alpha^b))$, $\forall \alpha^b$,

$$\frac{d\psi^a(\alpha^b)}{d\alpha^b} = \frac{d(l^a)^{-1}(u)}{d\alpha^b} = \frac{d(l^a)^{-1}(u)}{du} \frac{du}{d\alpha^b} = \frac{1}{(l^a)'((l^a)^{-1}(u))} \frac{du}{d\alpha^b} = -((l^a)^{-1}(u))^2 \frac{du}{d\alpha^b}.$$

Because $(l^a)^{-1}(u) = \psi^a(\alpha^b) \in [0, 1]$, $-((l^a)^{-1}(u))^2 \in [-1, 0]$. Moreover, because of the condition $|\frac{dh^a(\theta^a(\alpha^a, \alpha^b))}{d\alpha^b}| < 1$, we have

$$\left| \frac{d\psi^a(\alpha^b)}{d\alpha^b} \right| < 1.$$

Similarly, we can show that $|\frac{d\psi^b(\alpha^a)}{d\alpha^a}| < 1$ holds $\forall \alpha^a$ if $|\frac{\partial h^b(\theta^b(\alpha^a, \alpha^b))}{\partial \alpha^a}| < 1$. Therefore, $\mathcal{C}_1, \mathcal{C}_2$ have only one intersection, the equilibrium $(\hat{\alpha}^a, \hat{\alpha}^b)$ is unique.

Now consider the case when Condition 1(B) is satisfied.

Because $\frac{dl^s(\alpha^s)}{d\alpha^s} = -\frac{1}{(\alpha^s)^2} < -1, \forall \alpha^s \in (0, 1)$, and $-1 \leq \frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^s} \leq 0$ for any fixed $\alpha^{-s} \in [0, 1]$. Strictly decreasing function $l^s(\alpha^s)$ and non-increasing function $h^s(\theta^s(\alpha^a, \alpha^b))$ has exactly one intersection. Therefore, $\forall \alpha^b$, balanced set $\Psi^a(\alpha^b) = \{\bar{\alpha}^a : h^a(\theta^a(\bar{\alpha}^a, \alpha^b)) = l^a(\bar{\alpha}^a)\}$ has only one element, and they constitute continuous function $\bar{\alpha}^a = \psi^a(\alpha^b)$ (balanced function). Similarly, $\forall \alpha^b$, set $\Psi^a(\alpha^b) = \{\bar{\alpha}^a : h^a(\theta^a(\bar{\alpha}^a, \alpha^b)) = l^a(\bar{\alpha}^a)\}$ also has only one element, which forms continuous function $\bar{\alpha}^a = \psi^a(\alpha^b)$.

Because given any α^a , $h^a(\theta^a(\alpha^a, \alpha^b))$ is non-increasing in α^b . As α^b increases, the intersection with $l^a(\bar{\alpha}^a)$ is non-decreasing. Therefore, $\psi^a(\alpha^b)$ is non-decreasing in α^b . Similarly, $\psi^b(\alpha^a)$ is also non-decreasing in α^a .

On the 2D plane $\{(\alpha^a, \alpha^b) : \alpha^a \in [0, 1], \alpha^b \in [0, 1]\}$, two curves $\mathcal{C}_1 = \{(\alpha^a, \alpha^b) : \alpha^a = \psi^a(\alpha^b), \alpha^b \in [0, 1]\}$ and $\mathcal{C}_2 = \{(\alpha^a, \alpha^b) : \alpha^b = \psi^b(\alpha^a), \alpha^a \in [0, 1]\}$ are both continuous and non-decreasing. One sufficient condition to guarantee \mathcal{C}_1 and \mathcal{C}_2 have exact one intersection, is that $\frac{d\psi^a(\alpha^b)}{d\alpha^b} < 1, \forall \alpha^b \in [0, 1]$ and $\frac{d\psi^b(\alpha^a)}{d\alpha^a} < 1, \forall \alpha^a \in [0, 1]$. Using the same analysis as the case under Condition 1(A), we can show these sufficient conditions will hold if $|\frac{\partial h^a(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^b}| < 1$ and $|\frac{\partial h^b(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^a}| < 1, \forall \alpha^a, \alpha^b$.

Therefore, $\mathcal{C}_1, \mathcal{C}_2$ have only one intersection, the equilibrium $(\hat{\alpha}^a, \hat{\alpha}^b)$ is unique. \square

The proof of Corollary 1

Corollary 1. For any feature distribution $\{G_y^s(x)\}_{s,y}$, suppose that $|\frac{\partial G_y^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^u}| \leq M_y$ holds for some constant $M_y \in [0, \infty), \forall y \in \{0, 1\}, \forall u \in \{a, b\}$. Under either Condition 1(A) or 1(B), $\exists \epsilon_y^s > 0$ such that for any transitions that satisfy $|T_{y1}^s - T_{y0}^s| < \epsilon_y^s, s \in \{a, b\}, y \in \{0, 1\}$, the corresponding dynamics system has a unique equilibrium.

Proof. Define notations $\mathbb{G}_y^s := \mathbb{G}_y^s(\theta^s(\alpha^a, \alpha^b)), \Delta T_0^s := T_{01}^s - T_{00}^s$ and $\Delta T_1^s := T_{11}^s - T_{10}^s$.

$$h^s(\theta^s(\alpha^a, \alpha^b)) = \frac{(1 - T_{10}^s)\mathbb{G}_1^s + (1 - T_{11}^s)(1 - \mathbb{G}_1^s)}{T_{00}^s\mathbb{G}_0^s + T_{01}^s(1 - \mathbb{G}_0^s)} = \frac{(1 - T_{11}^s) + \Delta T_1^s\mathbb{G}_1^s}{T_{00}^s + \Delta T_0^s(1 - \mathbb{G}_0^s)}$$

Take derivative w.r.t. $\alpha^u, \forall u \in \{a, b\}$,

$$\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^u} = \frac{\Delta T_1^s \frac{\partial \mathbb{G}_1^s}{\partial \alpha^u} (T_{00}^s + \Delta T_0^s(1 - \mathbb{G}_0^s)) + \Delta T_0^s \frac{\partial \mathbb{G}_0^s}{\partial \alpha^u} ((1 - T_{11}^s) + \Delta T_1^s \mathbb{G}_1^s)}{(T_{00}^s + \Delta T_0^s(1 - \mathbb{G}_0^s))^2}$$

Consider case under Condition 1(A). Since $\Delta T_0^s < 0, \Delta T_1^s < 0, T_{00}^s + \Delta T_0^s(1 - \mathbb{G}_0^s) > 0$, and $(1 - T_{11}^s) + \Delta T_1^s\mathbb{G}_1^s > 0$, we have $|\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^u}| \leq |\frac{\Delta T_1^s M_1 T_{00}^s + \Delta T_0^s M_0(1 - T_{11}^s)}{(T_{01}^s)^2}|$.

Take $\epsilon_1^s = \epsilon_0^s = \frac{(T_{01}^s)^2}{M_1 T_{00}^s + M_0(1 - T_{11}^s)}$, if $|\Delta T_1^s| < \epsilon_1^s$ and $|\Delta T_0^s| < \epsilon_0^s$, then $|\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^u}| < 1$ holds. From Theorem 3, the equilibrium of dynamics 4 is unique.

Consider case under Condition 1(B).

Since $\Delta T_0^s > 0$ and $\Delta T_1^s > 0$, we have $|\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^u}| \leq \frac{\Delta T_1^s M_1 T_{01}^s + \Delta T_0^s M_0(1 - T_{10}^s)}{(T_{00}^s)^2}$.

Take $\epsilon_1^s = \epsilon_0^s = \frac{(T_{00}^s)^2}{M_1 T_{01}^s + M_0(1 - T_{10}^s)}$, if $\Delta T_1^s < \epsilon_1^s$ and $\Delta T_0^s < \epsilon_0^s$, then $|\frac{\partial h^s(\theta^s(\alpha^a, \alpha^b))}{\partial \alpha^u}| < 1$ holds. From Theorem 3, the equilibrium of dynamics 4 is unique. \square

The proof of Theorem 4.

Proof. $\forall s \in \{a, b\}$, an equilibrium $\hat{\alpha}_{\text{UN}}^s$ satisfies:

$$\frac{1 - g^{1s}(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))}{g^{0s}(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))} = \frac{1 - (T_{11}^s(1 - \mathbb{G}_1^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))) + T_{10}^s\mathbb{G}_1^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s)))}{T_{01}^s(1 - \mathbb{G}_0^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))) + T_{00}^s\mathbb{G}_0^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))} = \frac{1}{\hat{\alpha}_{\text{UN}}^s} - 1.$$

One solution to the above equation is:

$$\hat{\alpha}_{\text{UN}}^s = T_{11}^s(1 - \mathbb{G}_1^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))) + T_{10}^s\mathbb{G}_1^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s)) = T_{01}^s(1 - \mathbb{G}_0^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))) + T_{00}^s\mathbb{G}_0^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))$$

It shows that $\hat{\alpha}_{\text{UN}}^s$ is a convex combination of T_{00}^s, T_{01}^s , and also a convex combination of T_{10}^s, T_{11}^s .

$\forall \alpha_{\text{UN}}$ and $\mathbb{G}_0^s(x), \mathbb{G}_1^s(x)$, there is a set of transitions with $T_{00}^s < \alpha_{\text{UN}} < T_{01}^s$ and $T_{10}^s < \alpha_{\text{UN}} < T_{11}^s$ (satisfy Condition 1(B)), or $T_{01}^s < \alpha_{\text{UN}} < T_{00}^s$ and $T_{11}^s < \alpha_{\text{UN}} < T_{10}^s$ (satisfy Condition 1(A)), such that the above equation holds with $\hat{\alpha}_{\text{UN}}^s = \alpha_{\text{UN}}, \forall s \in \{a, b\}$, i.e., equitable equilibrium is attained.

Next we show that if $G_y^a(x) \neq G_y^b(x)$, then $\hat{\alpha}_{\mathcal{C}}^b \neq \hat{\alpha}_{\mathcal{C}}^a$ under these sets of transitions. Under the conditions of Theorem 3, $(\hat{\alpha}_{\mathcal{C}}^a, \hat{\alpha}_{\mathcal{C}}^b)$ is the intersection of two curves $\mathcal{C}_1 = \{(\alpha^a, \alpha^b) : \alpha^a =$

$\psi_C^a(\alpha^b), \alpha^b \in [0, 1]$ and $\mathcal{C}_2 = \{(\alpha^a, \alpha^b) : \alpha^b = \psi_C^b(\alpha^a), \alpha^a \in [0, 1]\}$; furthermore, let $\tilde{\alpha}_C^a, \tilde{\alpha}_C^b$ be defined such that $\tilde{\alpha}_C^a = \psi_C^a(\tilde{\alpha}_C^a), \tilde{\alpha}_C^b = \psi_C^b(\tilde{\alpha}_C^b)$, which are the intersections of $\alpha^a = \psi_C^a(\alpha^b)$ and $\alpha^a = \alpha^b$, as well as $\alpha^b = \psi_C^b(\alpha^a)$ and $\alpha^a = \alpha^b$, respectively. Then in order to prove $\hat{\alpha}_C^b \neq \hat{\alpha}_C^a$, it is sufficient to show $\tilde{\alpha}_C^a \neq \tilde{\alpha}_C^b$.

Given $\alpha^a = \alpha^b = \alpha_{\text{UN}}$, because $G_y^a(x) \neq G_y^b(x)$, we have $\theta_{\text{UN}}^s(\alpha_{\text{UN}}) \neq \theta_C^s(\alpha_{\text{UN}}, \alpha_{\text{UN}})$ and to satisfy Eqn. (3), there are only two possibilities: (1) $\theta_{\text{UN}}^a(\alpha_{\text{UN}}) > \theta_C^a(\alpha_{\text{UN}}, \alpha_{\text{UN}}), \theta_{\text{UN}}^b(\alpha_{\text{UN}}) < \theta_C^b(\alpha_{\text{UN}}, \alpha_{\text{UN}})$; (2) $\theta_{\text{UN}}^a(\alpha_{\text{UN}}) < \theta_C^a(\alpha_{\text{UN}}, \alpha_{\text{UN}}), \theta_{\text{UN}}^b(\alpha_{\text{UN}}) > \theta_C^b(\alpha_{\text{UN}}, \alpha_{\text{UN}})$.

WLOG, suppose the first case holds. Under Condition 1(B),

$$\frac{1 - g^{1b}(\theta_{\text{UN}}^b(\alpha_{\text{UN}}))}{g^{0b}(\theta_{\text{UN}}^b(\alpha_{\text{UN}}))} < \frac{1 - g^{1b}(\theta_C^b(\alpha_{\text{UN}}, \alpha_{\text{UN}}))}{g^{0b}(\theta_C^b(\alpha_{\text{UN}}, \alpha_{\text{UN}}))}; \quad \frac{1 - g^{1a}(\theta_{\text{UN}}^a(\alpha_{\text{UN}}))}{g^{0a}(\theta_{\text{UN}}^a(\alpha_{\text{UN}}))} > \frac{1 - g^{1a}(\theta_C^a(\alpha_{\text{UN}}, \alpha_{\text{UN}}))}{g^{0a}(\theta_C^a(\alpha_{\text{UN}}, \alpha_{\text{UN}}))}$$

It implies that $\tilde{\alpha}_C^b < \hat{\alpha}_{\text{UN}}^b = \hat{\alpha}_{\text{UN}}^a < \tilde{\alpha}_C^a$. Similarly, under Condition 1(A), $\tilde{\alpha}_C^b > \hat{\alpha}_{\text{UN}}^b = \hat{\alpha}_{\text{UN}}^a > \tilde{\alpha}_C^a$. Therefore, $\hat{\alpha}_C^a \neq \hat{\alpha}_C^b$.

In contrast, if $G_y^a(x) = G_y^b(x)$, we have $\theta_{\text{UN}}^s(\alpha) = \theta_C^s(\alpha, \alpha)$ and $\tilde{\alpha}_C^b = \tilde{\alpha}_C^a$. Therefore, $\hat{\alpha}_C^a = \hat{\alpha}_C^b$. \square

The proof of Theorem 5.

Proof. WLOG, suppose that $\hat{\alpha}_{\text{UN}}^a > \hat{\alpha}_{\text{UN}}^b$ in the proof. Let $\psi_C^a(\cdot), \psi_C^b(\cdot)$ be balanced functions as defined in Theorem 3 under constraint \mathcal{C} . Firstly, we show that $\hat{\alpha}_{\text{UN}}^b$ and $\hat{\alpha}_{\text{UN}}^a$ are solutions to $\begin{cases} \alpha^b = \psi_C^b(\alpha^a) \\ \alpha^a = \alpha^b \end{cases}$ and $\begin{cases} \alpha^a = \psi_C^a(\alpha^b) \\ \alpha^a = \alpha^b \end{cases}$, respectively, i.e., $\hat{\alpha}_{\text{UN}}^b = \psi_C^b(\hat{\alpha}_{\text{UN}}^b)$ and $\hat{\alpha}_{\text{UN}}^a = \psi_C^a(\hat{\alpha}_{\text{UN}}^a)$.

Because $G_y^a(x) = G_y^b(x), \forall y \in \{0, 1\}, \forall x$, when $\alpha^a = \alpha^b = \alpha$, we have $\gamma^a(x) = \gamma^b(x), \mathcal{P}_{\text{EqOpt}}^a(x) = \mathcal{P}_{\text{EqOpt}}^b(x)$ and $\mathcal{P}_{\text{DP}}^a(x) = \mathcal{P}_{\text{DP}}^b(x)$, which implies $\theta_C^a(\alpha, \alpha) = \theta_C^b(\alpha, \alpha)$; furthermore, the optimal fair policies of DP and EqOpt satisfy $\gamma^a(\theta_C^a(\alpha, \alpha)) = \gamma^b(\theta_C^b(\alpha, \alpha)) = \frac{u_-}{u_+ + u_-}$ according to the optimal fair policy equation:

$$\frac{p_a \alpha^a}{\gamma^a(\theta_{\text{EqOpt}}^a)} + \frac{p_b \alpha^b}{\gamma^b(\theta_{\text{EqOpt}}^b)} = \frac{p_a \alpha^a}{\frac{u_-}{u_+ + u_-}} + \frac{p_b \alpha^b}{\frac{u_-}{u_+ + u_-}}; \quad p_a \gamma^a(\theta_{\text{DP}}^a) + p_b \gamma^b(\theta_{\text{DP}}^b) = \frac{u_-}{u_+ + u_-}.$$

Because $\gamma^a(\theta_{\text{UN}}^a(\alpha)) = \gamma^b(\theta_{\text{UN}}^b(\alpha)) = \frac{u_-}{u_+ + u_-}$ we have $\gamma^a(\theta_{\text{UN}}^a(\alpha)) = \gamma^a(\theta_C^a(\alpha, \alpha)) = \gamma^b(\theta_{\text{UN}}^b(\alpha)) = \gamma^b(\theta_C^b(\alpha, \alpha))$ so that $\theta_C^a(\alpha, \alpha) = \theta_{\text{UN}}^a(\alpha) = \theta_C^b(\alpha, \alpha) = \theta_{\text{UN}}^b(\alpha)$ holds under any α . $\forall s \in \{a, b\}$, because $\hat{\alpha}_{\text{UN}}^s$ is the solution to balanced equation, i.e., $l^s(\hat{\alpha}_{\text{UN}}^s) = h^s(\theta_{\text{UN}}^s(\hat{\alpha}_{\text{UN}}^s))$. We have $l^s(\hat{\alpha}_{\text{UN}}^s) = h^s(\theta_C^s(\hat{\alpha}_{\text{UN}}^s, \hat{\alpha}_{\text{UN}}^s))$, which further implies $\hat{\alpha}_{\text{UN}}^s = \psi_C^s(\hat{\alpha}_{\text{UN}}^s)$.

Under Condition 1(B), according to the proof of Theorem 3, we know that $0 \leq \frac{d\psi_C^b(\alpha^a)}{d\alpha^a} < 1$ and $0 \leq \frac{d\psi_C^a(\alpha^b)}{d\alpha^b} < 1$. Because $\hat{\alpha}_{\text{UN}}^b = \psi_C^b(\hat{\alpha}_{\text{UN}}^b) < \hat{\alpha}_{\text{UN}}^a = \psi_C^a(\hat{\alpha}_{\text{UN}}^a)$, we have $\hat{\alpha}_{\text{UN}}^b < \psi_C^b(\alpha^a) < \alpha^a, \forall \alpha^a \in [\hat{\alpha}_{\text{UN}}^b, \hat{\alpha}_{\text{UN}}^a]$. Similarly, we have $\alpha^b < \psi_C^a(\alpha^b) < \hat{\alpha}_{\text{UN}}^a, \forall \alpha^b \in [\hat{\alpha}_{\text{UN}}^b, \hat{\alpha}_{\text{UN}}^a]$. Therefore, after representing the two balanced functions as two curves $\mathcal{C}_1 = \{(\alpha^a, \alpha^b) : \alpha^a = \psi_C^a(\alpha^b), \alpha^b \in [0, 1]\}$ and $\mathcal{C}_2 = \{(\alpha^a, \alpha^b) : \alpha^b = \psi_C^b(\alpha^a), \alpha^a \in [0, 1]\}$ on the 2D plane $\{(\alpha^a, \alpha^b) : \alpha^a \in [0, 1], \alpha^b \in [0, 1]\}$, the intersection $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$ of \mathcal{C}_1 and \mathcal{C}_2 satisfies: 1) $\hat{\alpha}_C^a > \hat{\alpha}_C^b$; 2) $\hat{\alpha}_{\text{UN}}^b < \hat{\alpha}_C^a < \hat{\alpha}_{\text{UN}}^a$; 3) $\hat{\alpha}_{\text{UN}}^b < \hat{\alpha}_C^b < \hat{\alpha}_{\text{UN}}^a$. Therefore, $|\hat{\alpha}_C^a - \hat{\alpha}_C^b| \leq |\hat{\alpha}_{\text{UN}}^a - \hat{\alpha}_{\text{UN}}^b|$.

Under Condition 1(A), according to the proof of Theorem 3, we know that $-1 < \frac{d\psi_C^b(\alpha^a)}{d\alpha^a} \leq 0$ and $-1 < \frac{d\psi_C^a(\alpha^b)}{d\alpha^b} \leq 0$. Because $\hat{\alpha}_{\text{UN}}^b = \psi_C^b(\hat{\alpha}_{\text{UN}}^b) < \hat{\alpha}_{\text{UN}}^a = \psi_C^a(\hat{\alpha}_{\text{UN}}^a)$, we have $\psi_C^b(\alpha^a) < \hat{\alpha}_{\text{UN}}^b, \forall \alpha^a > \hat{\alpha}_{\text{UN}}^b$. Similarly, we have $\psi_C^a(\alpha^b) > \hat{\alpha}_{\text{UN}}^a, \forall \alpha^b < \hat{\alpha}_{\text{UN}}^a$. Due to the existence of equilibrium, the intersection $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$ of \mathcal{C}_1 and \mathcal{C}_2 must satisfy: 1) $\hat{\alpha}_C^a > \hat{\alpha}_C^b$; 2) $\hat{\alpha}_{\text{UN}}^a < \hat{\alpha}_C^a$; 3) $\hat{\alpha}_C^b < \hat{\alpha}_{\text{UN}}^b$. Therefore, $|\hat{\alpha}_C^a - \hat{\alpha}_C^b| \geq |\hat{\alpha}_{\text{UN}}^a - \hat{\alpha}_{\text{UN}}^b|$. \square

The proof of Theorem 6.

Proof. The proof is under the conditions of Theorem 3 such that there is unique equilibrium of qualification rate. Under fairness constraint $\mathcal{C} = \text{EqOpt}$ or DP, consider 2D plane $\{(\alpha^a, \alpha^b) :$

$\alpha^a \in [0, 1], \alpha^b \in [0, 1]$, and note that equilibrium $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$ is the intersection of two curves $\mathcal{C}_1 = \{(\alpha^a, \alpha^b) : \alpha^a = \psi_C^a(\alpha^b), \alpha^b \in [0, 1]\}$ and $\mathcal{C}_2 = \{(\alpha^a, \alpha^b) : \alpha^b = \psi_C^b(\alpha^a), \alpha^a \in [0, 1]\}$. Consider a line $\{(\alpha^a, \alpha^b) : \alpha^a = \alpha^b, \alpha^a \in [0, 1], \alpha^b \in [0, 1]\}$, which has unique intersection $\tilde{\alpha}_C^a$ with \mathcal{C}_1 , and unique intersection $\tilde{\alpha}_C^b$ with \mathcal{C}_2 . That is, $\tilde{\alpha}_C^a = \psi_C^a(\tilde{\alpha}_C^a)$, $\tilde{\alpha}_C^b = \psi_C^b(\tilde{\alpha}_C^b)$.

First of all, we show that if $\frac{u_+}{u_-} \geq \frac{1-T_{10}}{T_{00}} \beta(\hat{x})$, under Condition 1(B), $\hat{\alpha}_{UN}^b < \hat{\alpha}_{UN}^a$.

By Condition 2, given any $\alpha^a = \alpha^b = \alpha$, the corresponding qualification profiles of $\mathcal{G}_a, \mathcal{G}_b$ satisfy the followings: $\gamma^b(\hat{x}) = \gamma^a(\hat{x})$; $\gamma^b(x) < \gamma^a(x), \forall x < \hat{x}$; $\gamma^b(x) > \gamma^a(x), \forall x > \hat{x}$. Let $\bar{\alpha}$ be qualification rate such that $\gamma^a(\hat{x}) = \gamma^b(\hat{x}) = \frac{u_-}{u_+ + u_-} \implies \frac{u_+}{u_-} = \beta(\hat{x})(\frac{1}{\bar{\alpha}} - 1)$, where $\beta(\hat{x}) := \frac{G_0^a(\hat{x})}{G_1^a(\hat{x})} = \frac{G_0^b(\hat{x})}{G_1^b(\hat{x})}$, then $\forall \alpha \in [\bar{\alpha}, 1]$, $\gamma^a(\theta_{UN}^a(\alpha)) = \gamma^b(\theta_{UN}^b(\alpha)) = \frac{u_-}{u_+ + u_-} < \frac{1}{\beta(\hat{x})(\frac{1}{\bar{\alpha}} - 1) + 1} = \gamma^a(\hat{x}) = \gamma^b(\hat{x})$. Thus, $\forall \alpha \in [\bar{\alpha}, 1]$, $\theta_{UN}^a(\alpha) < \theta_{UN}^b(\alpha) < \hat{x}$, which implies $G_1^a(\theta_{UN}^a(\alpha)) < G_1^b(\theta_{UN}^b(\alpha))$ and $G_0^a(\theta_{UN}^a(\alpha)) < G_0^b(\theta_{UN}^b(\alpha))$; furthermore, under Condition 1(B), we have

$$\frac{1 - T_{11}}{T_{01}} < \frac{1 - g^{1a}(\theta_{UN}^a(\alpha))}{g^{0a}(\theta_{UN}^a(\alpha))} < \frac{1 - g^{1b}(\theta_{UN}^b(\alpha))}{g^{0b}(\theta_{UN}^b(\alpha))} < \frac{1 - T_{10}}{T_{00}}, \forall \alpha \in [\bar{\alpha}, 1].$$

Because $\hat{\alpha}_{UN}^a$ and $\hat{\alpha}_{UN}^b$ are solutions to balance equations, i.e., $\frac{1}{\hat{\alpha}_{UN}^a} - 1 = \frac{1 - g^{1a}(\theta_{UN}^a(\hat{\alpha}_{UN}^a))}{g^{0a}(\theta_{UN}^a(\hat{\alpha}_{UN}^a))}$, $\frac{1}{\hat{\alpha}_{UN}^b} - 1 = \frac{1 - g^{1b}(\theta_{UN}^b(\hat{\alpha}_{UN}^b))}{g^{0b}(\theta_{UN}^b(\hat{\alpha}_{UN}^b))}$. If $\bar{\alpha} \leq \hat{\alpha}_{UN}^b$, the $\hat{\alpha}_{UN}^b < \hat{\alpha}_{UN}^a$ must hold under Condition 1(B). Next, we show that a sufficient condition of $\bar{\alpha} \leq \hat{\alpha}_{UN}^b$ is $\frac{u_+}{u_-} \geq \frac{1-T_{10}}{T_{00}} \beta(\hat{x})$.

$\frac{u_+}{u_-} \geq \frac{1-T_{10}}{T_{00}} \beta(\hat{x}) \implies \frac{1}{\bar{\alpha}} - 1 \geq \frac{1-T_{10}}{T_{00}}$. Since $\frac{1}{\hat{\alpha}_{UN}^b} - 1 < \frac{1-T_{10}}{T_{00}}$, we have $\frac{1}{\hat{\alpha}_{UN}^b} - 1 < \frac{1}{\bar{\alpha}} - 1$. Thus, $\bar{\alpha} \leq \hat{\alpha}_{UN}^b$. Therefore, if $\frac{u_+}{u_-} \geq \frac{1-T_{10}}{T_{00}} \beta(\hat{x})$, under Condition 1(B), $\hat{\alpha}_{UN}^b < \hat{\alpha}_{UN}^a$.

Fairness constraint Eq0pt. Secondly, we show that for Eq0pt fair policy, if $\frac{u_+}{u_-} \geq \frac{1-T_{10}}{T_{00}} \beta(\hat{x})$, under Condition 1(B), $\hat{\alpha}_{UN}^a - \hat{\alpha}_{UN}^b > \hat{\alpha}_{Eq0pt}^a - \hat{\alpha}_{Eq0pt}^b \geq 0$. Because two curves $\mathcal{C}_1, \mathcal{C}_2$ are monotonic increasing. It's sufficient to show two parts: (1) $\tilde{\alpha}_{Eq0pt}^a < \hat{\alpha}_{UN}^a, \tilde{\alpha}_{Eq0pt}^b > \hat{\alpha}_{UN}^b$; (2) $\tilde{\alpha}_{Eq0pt}^a \geq \tilde{\alpha}_{Eq0pt}^b$.

Under Eq0pt constraint, $\forall \alpha^a, \alpha^b, G_1^a(\theta_{Eq0pt}^a(\alpha^a, \alpha^b)) = G_1^b(\theta_{Eq0pt}^b(\alpha^a, \alpha^b))$ must hold so that $\theta_{Eq0pt}^a(\alpha^a, \alpha^b) = \theta_{Eq0pt}^b(\alpha^a, \alpha^b)$. Consider the case $\alpha^a = \alpha^b = \alpha, \forall \alpha \geq \bar{\alpha}$, we have $\theta_{Eq0pt}^a(\alpha, \alpha) = \theta_{Eq0pt}^b(\alpha, \alpha)$ and $\theta_{UN}^a(\alpha) < \theta_{UN}^b(\alpha)$. It implies that $\theta_{UN}^a(\alpha) < \theta_{Eq0pt}^a(\alpha, \alpha) = \theta_{Eq0pt}^b(\alpha, \alpha) < \theta_{UN}^b(\alpha) < \hat{x}$, otherwise Equation (3) will be violated. Therefore, the followings hold $\forall \alpha \in [\bar{\alpha}, 1]$,

$$\frac{1 - g^{1a}(\theta_{Eq0pt}^a(\alpha, \alpha))}{g^{0a}(\theta_{Eq0pt}^a(\alpha, \alpha))} > \frac{1 - g^{1a}(\theta_{UN}^a(\alpha))}{g^{0a}(\theta_{UN}^a(\alpha))}; \quad \frac{1 - g^{1b}(\theta_{Eq0pt}^b(\alpha, \alpha))}{g^{0b}(\theta_{Eq0pt}^b(\alpha, \alpha))} < \frac{1 - g^{1b}(\theta_{UN}^b(\alpha))}{g^{0b}(\theta_{UN}^b(\alpha))}.$$

$\forall s \in \{a, b\}$, $\tilde{\alpha}_{Eq0pt}^s$ is the solution to $\frac{1 - g^{1s}(\theta_{Eq0pt}^s(\alpha, \alpha))}{g^{0s}(\theta_{Eq0pt}^s(\alpha, \alpha))} = \frac{1}{\alpha} - 1$ while $\hat{\alpha}_{UN}^s$ is the solution to $\frac{1 - g^{1s}(\theta_{UN}^s(\alpha))}{g^{0s}(\theta_{UN}^s(\alpha))} = \frac{1}{\alpha} - 1$. Since $\bar{\alpha} \leq \hat{\alpha}_{UN}^b < \hat{\alpha}_{UN}^a$, it implies $\tilde{\alpha}_{Eq0pt}^a < \hat{\alpha}_{UN}^a, \tilde{\alpha}_{Eq0pt}^b > \hat{\alpha}_{UN}^b$.

Next, show that $\tilde{\alpha}_{Eq0pt}^a \geq \tilde{\alpha}_{Eq0pt}^b, \forall \alpha \geq \bar{\alpha}, \theta_{Eq0pt}^a(\alpha, \alpha) = \theta_{Eq0pt}^b(\alpha, \alpha)$ implies $G_1^a(\theta_{Eq0pt}^a(\alpha, \alpha)) = G_1^b(\theta_{Eq0pt}^b(\alpha, \alpha))$ and $G_0^a(\theta_{Eq0pt}^a(\alpha, \alpha)) \leq G_0^b(\theta_{Eq0pt}^b(\alpha, \alpha))$. Therefore,

$$\frac{1 - g^{1a}(\theta_{Eq0pt}^a(\alpha, \alpha))}{g^{0a}(\theta_{Eq0pt}^a(\alpha, \alpha))} \leq \frac{1 - g^{1b}(\theta_{Eq0pt}^b(\alpha, \alpha))}{g^{0b}(\theta_{Eq0pt}^b(\alpha, \alpha))}.$$

Intersections with function $\frac{1}{\alpha} - 1$ satisfies $\tilde{\alpha}_{Eq0pt}^a \geq \tilde{\alpha}_{Eq0pt}^b$.

It thus concludes that $\hat{\alpha}_{UN}^a - \hat{\alpha}_{UN}^b > \hat{\alpha}_{Eq0pt}^a - \hat{\alpha}_{Eq0pt}^b \geq 0$.

Fairness constraint DP. Finally, consider DP fair policy, where $\forall \alpha^a, \alpha^b, (1 - \alpha^a)G_0^a(\theta_{DP}^a(\alpha^a, \alpha^b)) + \alpha^a G_1^a(\theta_{DP}^a(\alpha^a, \alpha^b)) = (1 - \alpha^b)G_0^b(\theta_{DP}^b(\alpha^a, \alpha^b)) + \alpha^b G_1^b(\theta_{DP}^b(\alpha^a, \alpha^b))$ must hold.

We first show that under Condition 1(B), $\tilde{\alpha}_{DP}^a < \hat{\alpha}_{UN}^a, \tilde{\alpha}_{DP}^b > \hat{\alpha}_{UN}^b$. Consider the case $\alpha^a = \alpha^b = \alpha, \forall \alpha \geq \bar{\alpha}$. Since $\forall x, (1 - \alpha)G_0^a(x) + \alpha G_1^a(x) \geq (1 - \alpha)G_0^b(x) + \alpha G_1^b(x)$, $(1 - \alpha)G_0^a(\theta_{DP}^a(\alpha, \alpha)) +$

$\alpha \mathbb{G}_1^a(\theta_{\text{DP}}^a(\alpha, \alpha)) = (1 - \alpha) \mathbb{G}_0^b(\theta_{\text{DP}}^b(\alpha, \alpha)) + \alpha \mathbb{G}_1^b(\theta_{\text{DP}}^b(\alpha, \alpha))$ implies $\theta_{\text{DP}}^a(\alpha, \alpha) \geq \theta_{\text{DP}}^b(\alpha, \alpha)$. Because $\theta_{\text{UN}}^a(\alpha) < \theta_{\text{UN}}^b(\alpha), \forall \alpha \geq \bar{\alpha}$. It implies that $\theta_{\text{DP}}^a(\alpha, \alpha) > \theta_{\text{UN}}^a(\alpha)$ and $\hat{x} > \theta_{\text{UN}}^b(\alpha) > \theta_{\text{DP}}^b(\alpha, \alpha)$ must hold. Therefore, $\forall \alpha \in [\bar{\alpha}, 1]$,

$$\frac{1 - g^{1a}(\theta_{\text{DP}}^a(\alpha, \alpha))}{g^{0a}(\theta_{\text{DP}}^a(\alpha, \alpha))} > \frac{1 - g^{1a}(\theta_{\text{UN}}^a(\alpha))}{g^{0a}(\theta_{\text{UN}}^a(\alpha))}; \quad \frac{1 - g^{1b}(\theta_{\text{DP}}^b(\alpha, \alpha))}{g^{0b}(\theta_{\text{DP}}^b(\alpha, \alpha))} < \frac{1 - g^{1b}(\theta_{\text{UN}}^b(\alpha))}{g^{0b}(\theta_{\text{UN}}^b(\alpha))}$$

Similar to reasoning in Eq0pt case, we have $\tilde{\alpha}_{\text{DP}}^a < \hat{\alpha}_{\text{UN}}^a, \tilde{\alpha}_{\text{DP}}^b > \hat{\alpha}_{\text{UN}}^b$.

Different from Eq0pt fairness where $\tilde{\alpha}_{\text{Eq0pt}}^a \geq \tilde{\alpha}_{\text{Eq0pt}}^b$, both $\tilde{\alpha}_{\text{DP}}^a \geq \tilde{\alpha}_{\text{DP}}^b$ and $\tilde{\alpha}_{\text{DP}}^a \leq \tilde{\alpha}_{\text{DP}}^b$ are likely to occur, depending on distributions $G_0^a(x), G_0^b(x), G_1^a(x)$ and $G_1^b(x)$. It is because $\theta_{\text{DP}}^a(\alpha, \alpha) > \theta_{\text{DP}}^b(\alpha, \alpha)$ can result in either $\mathbb{G}_0^a(\theta_{\text{DP}}^a(\alpha, \alpha)) \leq \mathbb{G}_0^b(\theta_{\text{DP}}^b(\alpha, \alpha))$ or $\mathbb{G}_0^a(\theta_{\text{DP}}^a(\alpha, \alpha)) \geq \mathbb{G}_0^b(\theta_{\text{DP}}^b(\alpha, \alpha))$.

For these two outcomes, if $\tilde{\alpha}_{\text{DP}}^a \geq \tilde{\alpha}_{\text{DP}}^b$, then DP fair policy results in a more equitable equilibrium than unconstrained policy; if $\tilde{\alpha}_{\text{DP}}^a \leq \tilde{\alpha}_{\text{DP}}^b$, it means the disadvantaged group is flipped from \mathcal{G}_b to \mathcal{G}_a . \square

The proof of Proposition 1.

Proof. In the proof, we simplify the notations by removing subscript \mathcal{C} .

Let $\psi^s(\cdot), \psi^{s'}(\cdot)$ be balanced function of policies (θ^a, θ^b) and $(\theta^{a'}, \theta^{b'})$, respectively.

According to the balanced equation (5),

$$\frac{1}{\alpha^s} - 1 = \frac{1 - g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))} = \frac{1 - (T_{11}^s(1 - \mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b))) + T_{10}^s \mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b)))}{T_{01}^s(1 - \mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b))) + T_{00}^s \mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b))}.$$

Under Condition (B), $\forall \alpha^a, \alpha^b \in [0, 1], \theta^{a'}(\alpha^a, \alpha^b) < \theta^a(\alpha^a, \alpha^b)$ and $\theta^{b'}(\alpha^a, \alpha^b) < \theta^b(\alpha^a, \alpha^b)$.

Under Condition (A), $\forall \alpha^a, \alpha^b \in [0, 1], \theta^{a'}(\alpha^a, \alpha^b) > \theta^a(\alpha^a, \alpha^b)$ and $\theta^{b'}(\alpha^a, \alpha^b) > \theta^b(\alpha^a, \alpha^b)$.

Both imply that $\frac{1 - g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))} > \frac{1 - g^{1s}(\theta^{s'}(\alpha^a, \alpha^b))}{g^{0s}(\theta^{s'}(\alpha^a, \alpha^b))}$, and $\forall \alpha^a, \alpha^b \in [0, 1], \psi^a(\alpha^b) < \psi^{a'}(\alpha^b)$ and $\psi^b(\alpha^a) < \psi^{b'}(\alpha^a)$ hold. As a consequence, $\hat{\alpha}^{a'} > \hat{\alpha}^a$ and $\hat{\alpha}^{b'} > \hat{\alpha}^b$.

Now consider the long-run average utility of institute $\bar{U}(\theta^a, \theta^b) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathcal{U}_t(\theta^a, \theta^b)$, where the instantaneous utility at t under threshold policies θ^a, θ^b is

$$\begin{aligned} \mathcal{U}_t(\theta^a, \theta^b) &= \sum_{s=a,b} p_s \mathbb{E}_{X_t | S=s} [\mathbf{1}(X_t \geq \theta^s) (\gamma_t^s(X_t)(u_+ + u_-) - u_-)] \\ &= \sum_{s=a,b} p_s \int_{\theta^s}^{\infty} (\gamma_t^s(x)(u_+ + u_-) - u_-) \mathbb{P}(X_t = x | S = s) dx \\ &= \sum_{s=a,b} p_s \int_{\theta^s}^{\infty} \alpha_t^s (G_1^s(x)u_+ + G_0^s(x)u_-) - G_0^s(x)u_- dx \end{aligned}$$

In the followings, we use a special case ($\mathcal{C} = \text{Eq0pt}, G_y^a(x) = G_y^b(x), \forall x, y = 0, 1$, under Condition 1(B)) to show that $\bar{U}(\theta^{a'}, \theta^{b'}) > \bar{U}(\theta^a, \theta^b)$ can be attained, i.e., the long-run average utility under policy $(\theta^{a'}, \theta^{b'})$ can be higher than myopic optimal policy (θ^a, θ^b) .

Since the qualification rates of two groups converge to equilibrium, $\bar{U}(\theta^a, \theta^b) = \mathcal{U}_{\infty}(\theta^a, \theta^b)$ is the same as instantaneous expected utility of institute at the equilibrium state. To show that $\bar{U}(\theta^{a'}, \theta^{b'}) > \bar{U}(\theta^a, \theta^b)$, we prove the following holds,

$$\sum_{s=a,b} p_s \int_{\theta^{s'}(\hat{\alpha}^{a'}, \hat{\alpha}^{b'})}^{\infty} f(x; \hat{\alpha}^{s'}) dx > \sum_{s=a,b} p_s \int_{\theta^s(\hat{\alpha}^a, \hat{\alpha}^b)}^{\infty} f(x; \hat{\alpha}^s) dx \quad (7)$$

where $f(x; \hat{\alpha}^s) := \hat{\alpha}^s (G_1^s(x)u_+ + G_0^s(x)u_-) - G_0^s(x)u_-$.

Because $\hat{\alpha}^{s'} > \hat{\alpha}^s$, $\theta^{s'}(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}) < \theta^s(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}) < \theta^s(\hat{\alpha}^a, \hat{\alpha}^b)$ holds under Condition (B). LHS of above inequality can be written as

$$\sum_{s=a,b} p_s \left(\int_{\theta^{s'}(\hat{\alpha}^{a'}, \hat{\alpha}^{b'})}^{\theta^s(\hat{\alpha}^a, \hat{\alpha}^b)} f(x; \hat{\alpha}^{s'}) dx + \int_{\theta^s(\hat{\alpha}^a, \hat{\alpha}^b)}^{\infty} f(x; \hat{\alpha}^{s'}) dx \right).$$

Inequality (7) can further be re-organized,

$$\sum_{s=a,b} p_s \int_{\theta^{s'}(\hat{\alpha}^{a'}, \hat{\alpha}^{b'})}^{\theta^s(\hat{\alpha}^a, \hat{\alpha}^b)} f(x; \hat{\alpha}^{s'}) dx > \sum_{s=a,b} p_s \int_{\theta^s(\hat{\alpha}^a, \hat{\alpha}^b)}^{\infty} (f(x; \hat{\alpha}^s) - f(x; \hat{\alpha}^{s'})) dx \quad (8)$$

Consider a special case where $\mathcal{C} = \text{Eq0pt}$ and $G_y^a(x) = G_y^b(x) = G_y(x), \forall x, \forall y \in \{0, 1\}$. Then $\forall \alpha^a, \alpha^b$, we have $\theta^a(\alpha^a, \alpha^b) = \theta^b(\alpha^a, \alpha^b)$ and $\theta^{a'}(\alpha^a, \alpha^b) = \theta^{b'}(\alpha^a, \alpha^b)$. Inequality (8) can be reduced to the following, $\forall s \in \{a, b\}$, simplify notations and let $\hat{\theta} := \theta^s(\hat{\alpha}^a, \hat{\alpha}^b)$, $\hat{\theta}' := \theta^{s'}(\hat{\alpha}^{a'}, \hat{\alpha}^{b'})$.

$$\begin{aligned} & (p_a \hat{\alpha}^{a'} + p_b \hat{\alpha}^{b'}) (\mathbb{G}_1(\hat{\theta}) - \mathbb{G}_1(\hat{\theta}')) u_+ \\ & + \underbrace{\left(u_+ (1 - \mathbb{G}_1(\hat{\theta})) + u_- (1 - \mathbb{G}_0(\hat{\theta})) \right) (p_a (\hat{\alpha}^{a'} - \hat{\alpha}^a) + p_b (\hat{\alpha}^{b'} - \hat{\alpha}^b))}_{\text{term 1}} \\ & > (p_a (1 - \hat{\alpha}^{a'}) + p_b (1 - \hat{\alpha}^{b'})) (\mathbb{G}_0(\hat{\theta}) - \mathbb{G}_0(\hat{\theta}')) u_- \end{aligned} \quad (9)$$

Because $\frac{1}{\hat{\alpha}^{s'}} - 1 = \frac{1 - g^{1s}(\hat{\theta}')}{g^{0s}(\hat{\theta}')}$ and $\frac{1}{\hat{\alpha}^s} - 1 = \frac{1 - g^{1s}(\hat{\theta})}{g^{0s}(\hat{\theta})}$.

$$\hat{\alpha}^{s'} - \hat{\alpha}^s > \frac{T_{01}^s - T_{00}^s}{1 - T_{10}^s + T_{01}^s} (\mathbb{G}_0(\hat{\theta}) - \mathbb{G}_0(\hat{\theta}'))$$

We have **term 1** >

$$\underbrace{\left(\frac{u_+}{u_-} (1 - \mathbb{G}_1(\hat{\theta})) + (1 - \mathbb{G}_0(\hat{\theta})) \right) \left(p_a \frac{T_{01}^a - T_{00}^a}{1 - T_{10}^a + T_{01}^a} + p_b \frac{T_{01}^b - T_{00}^b}{1 - T_{10}^b + T_{01}^b} \right)}_{:=h(\hat{\theta}) > 0} (\mathbb{G}_0(\hat{\theta}) - \mathbb{G}_0(\hat{\theta}')) u_-$$

For the optimal Eq0pt fair threshold $\theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'})$, the following holds

$$\begin{aligned} & (p_a \hat{\alpha}^{a'} + p_b \hat{\alpha}^{b'}) G_1(\theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'})) u_+ = (p_a (1 - \hat{\alpha}^{a'}) + p_b (1 - \hat{\alpha}^{b'})) G_0(\theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'})) u_- \\ & (p_a \hat{\alpha}^{a'} + p_b \hat{\alpha}^{b'}) G_1(x) u_+ > (p_a (1 - \hat{\alpha}^{a'}) + p_b (1 - \hat{\alpha}^{b'})) G_0(x) u_-, \forall x > \theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}) \\ & (p_a \hat{\alpha}^{a'} + p_b \hat{\alpha}^{b'}) G_1(x) u_+ < (p_a (1 - \hat{\alpha}^{a'}) + p_b (1 - \hat{\alpha}^{b'})) G_0(x) u_-, \forall x < \theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}) \end{aligned}$$

It implies that \exists some $\delta > 0$ s.t. $\forall x \in (\theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}) - \delta, \theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}) + \delta) := \mathcal{B}(\theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}), \delta)$,

$$(p_a \hat{\alpha}^{a'} + p_b \hat{\alpha}^{b'}) G_1(x) u_+ + h(\hat{\theta}) G_0(x) u_- > (p_a (1 - \hat{\alpha}^{a'}) + p_b (1 - \hat{\alpha}^{b'})) G_0(x) u_-.$$

$\hat{\theta}, \hat{\theta}' \in \mathcal{B}(\theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}), \delta)$ can be satisfied as long as $|\theta^s(\alpha^a, \alpha^b) - \theta^{s'}(\alpha^a, \alpha^b)| \leq \epsilon$ for some sufficiently small $\epsilon > 0$.

Using the mean value theorem, $\exists G_y(x)$ and $\tilde{x} \in (\hat{\theta}', \hat{\theta}) \subset \mathcal{B}(\theta(\hat{\alpha}^{a'}, \hat{\alpha}^{b'}), \delta)$ s.t.

$$\begin{aligned} & (p_a \hat{\alpha}^{a'} + p_b \hat{\alpha}^{b'}) (\mathbb{G}_1(\hat{\theta}) - \mathbb{G}_1(\hat{\theta}')) u_+ + h(\hat{\theta}) (\mathbb{G}_0(\hat{\theta}) - \mathbb{G}_0(\hat{\theta}')) u_- \\ & = \left((p_a \hat{\alpha}^{a'} + p_b \hat{\alpha}^{b'}) G_1(\tilde{x}) u_+ + h(\hat{\theta}) G_0(\tilde{x}) u_- \right) (\hat{\theta} - \hat{\theta}') \\ & > \left((p_a (1 - \hat{\alpha}^{a'}) + p_b (1 - \hat{\alpha}^{b'})) G_0(\tilde{x}) u_- \right) (\hat{\theta} - \hat{\theta}') \\ & \geq (p_a (1 - \hat{\alpha}^{a'}) + p_b (1 - \hat{\alpha}^{b'})) (\mathbb{G}_0(\hat{\theta}) - \mathbb{G}_0(\hat{\theta}')) u_-. \end{aligned}$$

Therefore, inequality (9) holds and $\bar{U}(\theta^{a'}, \theta^{b'}) > \bar{U}(\theta^a, \theta^b)$. \square

The proof of Proposition 2 To ensure $\alpha_t^s \rightarrow \hat{\alpha}$, threshold policy $\theta^s(\alpha^s)$ as a function of $\alpha^s \in [0, 1]$ should be designed such that $\frac{1-g^{1s}(\theta^s(\alpha^s))}{g^{0s}(\theta^s(\alpha^s))} = \frac{1}{\alpha^s} - 1$ has a unique solution $\hat{\alpha}$. Let $\mathcal{I}'_s := \left[\frac{1-\max\{T_{11}^s, T_{10}^s\}}{\max\{T_{01}^s, T_{00}^s\}}, \frac{1-\min\{T_{11}^s, T_{10}^s\}}{\min\{T_{01}^s, T_{00}^s\}} \right]$, then $\frac{1-g^{1s}(\theta^s(\alpha^s))}{g^{0s}(\theta^s(\alpha^s))} \in \mathcal{I}'_s$ for any threshold policy $\theta^s(\alpha^s)$.

If $\mathcal{I}_a \cap \mathcal{I}_b = \emptyset$, then $\mathcal{I}'_a \cap \mathcal{I}'_b = \emptyset$ and $\frac{1-g^{1a}(\theta^a(\alpha))}{g^{0a}(\theta^a(\alpha))} = \frac{1-g^{1b}(\theta^b(\alpha))}{g^{0b}(\theta^b(\alpha))}$ can never be attained, i.e., no threshold policy can result in equitable equilibrium.

If $\mathcal{I}_a \cap \mathcal{I}_b \neq \emptyset$, then $\forall \hat{\alpha} \in \mathcal{I}_a \cap \mathcal{I}_b$ and $\forall s \in \{a, b\}$, there exists threshold policy $\theta^s(\alpha^s)$ such that $\frac{1-g^{1s}(\theta^s(\hat{\alpha}))}{g^{0s}(\theta^s(\hat{\alpha}))} = \frac{1}{\hat{\alpha}} - 1$. Specifically, under Condition 1(B) (resp. 1(A)), function

$$h^s(x) := \frac{1-g^{1s}(x)}{g^{0s}(x)} = \frac{1-(T_{11}^s(1-\mathbb{G}_1^s(x)) + T_{10}^s\mathbb{G}_1^s(x))}{T_{01}^s(1-\mathbb{G}_0^s(x)) + T_{00}^s\mathbb{G}_0^s(x)}$$

is strictly increasing (resp. decreasing) in $x \in (-\infty, +\infty)$ from $\frac{1-T_{11}^s}{T_{01}^s}$ (resp. $\frac{1-T_{10}^s}{T_{00}^s}$) to $\frac{1-T_{10}^s}{T_{00}^s}$ (resp. $\frac{1-T_{11}^s}{T_{01}^s}$) and any non-increasing function $\theta^s(\alpha^s)$ that satisfies $\theta^s(\hat{\alpha}) = (h^s)^{-1}(\frac{1}{\hat{\alpha}} - 1)$ can result in $\alpha_t^s \rightarrow \hat{\alpha}$, where $(h^s)^{-1}(\cdot)$ is the inverse function of $h^s(\cdot)$.

The proof of Proposition 3.

Proof. According to the balanced equation (5),

$$\frac{1}{\alpha^s} - 1 = \frac{1-g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))} = \frac{1-(T_{11}^s(1-\mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b))) + T_{10}^s\mathbb{G}_1^s(\theta^s(\alpha^a, \alpha^b)))}{T_{01}^s(1-\mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b))) + T_{00}^s\mathbb{G}_0^s(\theta^s(\alpha^a, \alpha^b))}.$$

$\forall \alpha^a, \alpha^b \in [0, 1]$, increasing any T_{yd}^s decreases $\frac{1-g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))}$. Let $\psi^{s'}(\cdot)$ be the consequent balanced function after increasing T_{yd}^s , and $\hat{\alpha}^{s'}$ be corresponding equilibrium. Given any $\alpha^a, \alpha^b \in [0, 1]$, we have $\psi^a(\alpha^b) < \psi^{a'}(\alpha^b)$ and $\psi^b(\alpha^a) < \psi^{b'}(\alpha^a)$. Therefore, $\hat{\alpha}^{a'} > \hat{\alpha}^a$ and $\hat{\alpha}^{b'} > \hat{\alpha}^b$. \square

The proof of Proposition 4.

Proof. Let $\bar{\theta}_C^s(\alpha^a, \alpha^b)$ be threshold under estimated qualification profile $\bar{\gamma}^s(x)$.

First consider UN policy.

Since $\bar{\gamma}_t^a(x) = \gamma_t^a(x)$, $\forall t, x$ and $\mathcal{G}_a, \mathcal{G}_b$ are decoupled, we have $\bar{\alpha}_{\text{UN}}^a = \hat{\alpha}_{\text{UN}}^a$.

For disadvantaged group \mathcal{G}_b , given any α^b , threshold $\bar{\theta}_{\text{UN}}^b(\alpha^b)$ is chosen such that $\bar{\gamma}^b(\bar{\theta}_{\text{UN}}^b(\alpha^b)) = \frac{u_-}{u_+ + u_-} = \gamma^b(\theta_{\text{UN}}^b(\alpha^b))$. Since $\bar{\gamma}_t^b(x) \leq \gamma_t^b(x)$, $\forall x$ and are non-decreasing, we have $\bar{\theta}_{\text{UN}}^b(\alpha^b) \geq \theta_{\text{UN}}^b(\alpha^b)$, i.e., the threshold chosen based on estimated qualification profile is higher so that the people from disadvantaged group is under-selected.

Under Condition 1(B), $\frac{1-g^{1b}(\bar{\theta}_{\text{UN}}^b(\alpha^b))}{g^{0b}(\bar{\theta}_{\text{UN}}^b(\alpha^b))} \geq \frac{1-g^{1b}(\theta_{\text{UN}}^b(\alpha^b))}{g^{0b}(\theta_{\text{UN}}^b(\alpha^b))}$, $\forall \alpha^b$, so that their intersections with $\frac{1}{\alpha^b} - 1$ satisfy: $\bar{\alpha}_{\text{UN}}^b \leq \hat{\alpha}_{\text{UN}}^b$. Similarly, under Condition 1(A), $\frac{1-g^{1b}(\bar{\theta}_{\text{UN}}^b(\alpha^b))}{g^{0b}(\bar{\theta}_{\text{UN}}^b(\alpha^b))} \leq \frac{1-g^{1b}(\theta_{\text{UN}}^b(\alpha^b))}{g^{0b}(\theta_{\text{UN}}^b(\alpha^b))}$, and $\bar{\alpha}_{\text{UN}}^b \geq \hat{\alpha}_{\text{UN}}^b$.

Now consider $\mathcal{C} = \text{DP}$ or EqOpt optimal fair policy.

Because $(\bar{\alpha}_C^a, \bar{\alpha}_C^b)$ and $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$ satisfy the following:

$$\begin{aligned} \frac{1}{\bar{\alpha}_C^b} - 1 &= \frac{1-g^{1,b}(\bar{\theta}_C^b(\bar{\alpha}_C^a, \bar{\alpha}_C^b))}{g^{0,b}(\bar{\theta}_C^b(\bar{\alpha}_C^a, \bar{\alpha}_C^b))}, & \frac{1}{\bar{\alpha}_C^a} - 1 &= \frac{1-g^{1,a}(\bar{\theta}_C^a(\bar{\alpha}_C^a, \bar{\alpha}_C^b))}{g^{0,a}(\bar{\theta}_C^a(\bar{\alpha}_C^a, \bar{\alpha}_C^b))}; \\ \frac{1}{\hat{\alpha}_C^b} - 1 &= \frac{1-g^{1,b}(\theta_C^b(\hat{\alpha}_C^a, \hat{\alpha}_C^b))}{g^{0,b}(\theta_C^b(\hat{\alpha}_C^a, \hat{\alpha}_C^b))}, & \frac{1}{\hat{\alpha}_C^a} - 1 &= \frac{1-g^{1,a}(\theta_C^a(\hat{\alpha}_C^a, \hat{\alpha}_C^b))}{g^{0,a}(\theta_C^a(\hat{\alpha}_C^a, \hat{\alpha}_C^b))} \end{aligned}$$

Given any fixed α^a , $\bar{\theta}_C^b(\alpha^a, \alpha^b) \geq \theta_C^b(\alpha^a, \alpha^b)$, $\forall \alpha^b$. Therefore, their intersections with function $\frac{1}{\alpha^b} - 1$ satisfy $\bar{\psi}_C^b(\alpha^a) \leq \psi_C^b(\alpha^a)$ under Condition 1(B); $\bar{\psi}_C^b(\alpha^a) \geq \psi_C^b(\alpha^a)$ under Condition 1(A).

Given any fixed α^b , $\bar{\theta}_C^a(\alpha^a, \alpha^b) \geq \theta_C^a(\alpha^a, \alpha^b)$, $\forall \alpha^a$. Although $\gamma_t^a(x)$ is true, $\bar{\theta}_C^a$ is still affected by \mathcal{G}_b under fairness constraint. Therefore, the intersections with function $\frac{1}{\alpha^a} - 1$ satisfy $\bar{\psi}_C^a(\alpha^b) \leq \psi_C^a(\alpha^b)$ under Condition 1(B); $\bar{\psi}_C^a(\alpha^b) \geq \psi_C^a(\alpha^b)$ under Condition 1(A).

Therefore, $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$ as the intersection of two curves $\alpha^b = \psi_C^b(\alpha^a)$ and $\alpha^a = \psi_C^a(\alpha^b)$, and $(\bar{\alpha}_C^a, \bar{\alpha}_C^b)$ as the intersection of two curves $\alpha^b = \bar{\psi}_C^a(\alpha^b)$ and $\alpha^a = \bar{\psi}_C^b(\alpha^a)$ has the relation: $\hat{\alpha}_C^a \geq \bar{\alpha}_C^a$ and $\hat{\alpha}_C^b \geq \bar{\alpha}_C^b$ under Condition 1(B); $\hat{\alpha}_C^a \leq \bar{\alpha}_C^a$ and $\hat{\alpha}_C^b \leq \bar{\alpha}_C^b$ under Condition 1(A). \square

The Proof of Theorem 7.

Proof. In the proof, let $G_{11}(x) = G_{10}(x) = G_1(x)$ and remove subscript A to simplify the notation, i.e., $G_{0d}^A(\theta) := G_{0d}(\theta)$, $\forall d \in \{0, 1\}$.

Under Assumption 3 and the case when $G_{11}(x) = G_{10}(x) = G_1(x)$, $G_{01}(x) \neq G_{00}(x)$, use the fact that $\zeta_t^{11} + \zeta_t^{10} = \alpha_t$, $\zeta_t^{11} + \zeta_t^{10} + \zeta_t^{01} + \zeta_t^{00} = 1$, dynamics of qualification rate can be simplified as follows, where θ_t is the threshold of the policy at t .

$$\begin{aligned}\alpha_{t+1} &= [T_{11}(1 - \mathbb{G}_1(\theta_t)) + T_{10}\mathbb{G}_1(\theta_t) - T_{01}(1 - \mathbb{G}_{01}(\theta_t)) - T_{00}\mathbb{G}_{01}(\theta_t)]\alpha_t \\ &\quad + T_{01}(1 - \mathbb{G}_{01}(\theta_t)) + T_{00}\mathbb{G}_{01}(\theta_t) + (\mathbb{G}_{01}(\theta_t) - \mathbb{G}_{00}(\theta_t))(T_{01} - T_{00})\zeta_t^{00} \\ \zeta_{t+1}^{00} &= [(1 - T_{10})\mathbb{G}_1(\theta_t) - (1 - T_{00})\mathbb{G}_{01}(\theta_t)]\alpha_t + (1 - T_{00})[\mathbb{G}_{01}(\theta_t)(1 - \zeta_t^{00}) + \mathbb{G}_{00}(\theta_t)\zeta_t^{00}]\end{aligned}$$

and qualification profile is reduced to $\gamma_t(x) = \frac{1}{1 + \frac{G_{00}(x)\zeta_t^{00} + G_{01}(x)\zeta_t^{01}}{G_1(x)\alpha_t}}$.

The scenario A has the equilibrium if there exists α , ζ^{00} such that the following holds:

$$\begin{aligned}\alpha &= [T_{11}(1 - \mathbb{G}_1(\theta)) + T_{10}\mathbb{G}_1(\theta) - T_{01}(1 - \mathbb{G}_{01}(\theta)) - T_{00}\mathbb{G}_{01}(\theta)]\alpha \\ &\quad + T_{01}(1 - \mathbb{G}_{01}(\theta)) + T_{00}\mathbb{G}_{01}(\theta) + (\mathbb{G}_{01}(\theta) - \mathbb{G}_{00}(\theta))(T_{01} - T_{00})\zeta^{00} \\ \zeta^{00} &= [(1 - T_{10})\mathbb{G}_1(\theta) - (1 - T_{00})\mathbb{G}_{01}(\theta)]\alpha + (1 - T_{00})[\mathbb{G}_{01}(\theta)(1 - \zeta^{00}) + \mathbb{G}_{00}(\theta)\zeta^{00}]\end{aligned}$$

Re-organize above equations,

$$\begin{aligned}\frac{1}{\alpha} - 1 &= \frac{1 - T_{11}(1 - \mathbb{G}_1(\theta)) - T_{10}\mathbb{G}_1(\theta) - (\mathbb{G}_{01}(\theta) - \mathbb{G}_{00}(\theta))(T_{01} - T_{00})\zeta^{00}}{T_{01}(1 - \mathbb{G}_{01}(\theta)) + T_{00}\mathbb{G}_{01}(\theta) + (\mathbb{G}_{01}(\theta) - \mathbb{G}_{00}(\theta))(T_{01} - T_{00})\zeta^{00}} = h_A(\theta_A(\zeta^{00}, \alpha)) \\ \zeta^{00} &= \frac{[(1 - T_{10})\mathbb{G}_1(\theta) - (1 - T_{00})\mathbb{G}_{01}(\theta)]\alpha + (1 - T_{00})\mathbb{G}_{01}(\theta)}{1 + (1 - T_{00})(\mathbb{G}_{01}(\theta) - \mathbb{G}_{00}(\theta))} = g_A(\theta_A(\zeta^{00}, \alpha))\end{aligned}$$

where $\theta := \theta_A(\zeta^{00}, \alpha)$ in equations for simplification. Given any ζ^{00} , $h_A(\theta_A(\zeta^{00}, \alpha))$ is continuous in α , there exists at least one $\bar{\alpha} := \psi(\zeta^{00})$ such that $\frac{1}{\bar{\alpha}} - 1 = h_A(\theta_A(\zeta^{00}, \bar{\alpha}))$ holds.

Given any $\alpha \in [0, 1]$, $g_A(\theta_A(\zeta^{00}, \alpha)) \in (0, 1)$ is continuous in ζ^{00} , while ζ^{00} is strictly increasing from 0 to 1. Therefore, there exists at least one $\bar{\zeta}^{00} := \eta(\alpha)$ such that $\bar{\zeta}^{00} = g_A(\theta_A(\bar{\zeta}^{00}, \alpha))$.

Consider a 2D plane $\{(\alpha, \zeta^{00}) : \zeta^{00} \in [0, 1], \alpha \in [0, 1]\}$, two sets of curves $\mathcal{C}_1 = \{(\alpha, \zeta^{00}) : \alpha = \psi(\zeta^{00}), \zeta^{00} \in (0, 1)\}$ and $\mathcal{C}_2 = \{(\alpha, \zeta^{00}) : \zeta^{00} = \eta(\alpha), \alpha \in (0, 1)\}$, then \mathcal{C}_1 and \mathcal{C}_2 must have at least one intersection, noted as $(\hat{\alpha}_A, \hat{\zeta}^{00})$. However, in order for $(\hat{\alpha}_A, \hat{\zeta}^{00})$ to be a valid equilibrium of the system for scenario A , it must satisfy $\hat{\alpha}_A + \hat{\zeta}^{00} \leq 1$. We show it in the following.

Sum up above balanced equations regarding α and ζ^{00} gives

$$\begin{aligned}\hat{\alpha}_A + \hat{\zeta}^{00} &= [T_{11} - T_{01} + (1 - T_{11})\mathbb{G}_1(\theta) + (T_{01} - 1)\mathbb{G}_{01}(\theta)]\hat{\alpha}_A \\ &\quad + (\mathbb{G}_{01}(\theta) - \mathbb{G}_{00}(\theta))(T_{01} - 1)\hat{\zeta}^{00} + T_{01} + (1 - T_{01})\mathbb{G}_{01}(\theta) \\ &= T_{01} + (1 - T_{01})\mathbb{G}_{01}(\theta)(1 - \hat{\alpha}_A - \hat{\zeta}^{00}) + (1 - T_{01})\mathbb{G}_{00}(\theta)\hat{\zeta}^{00} \\ &\quad + \hat{\alpha}_A(T_{11} - T_{01} + (1 - T_{11})\mathbb{G}_1(\theta)) \\ &\leq T_{01} + \max\{(1 - T_{01})\mathbb{G}_{01}(\theta), (1 - T_{01})\mathbb{G}_{00}(\theta), T_{11} - T_{01} + (1 - T_{11})\mathbb{G}_1(\theta)\}\end{aligned}$$

Since

$$\begin{aligned}T_{01} + (1 - T_{01})\mathbb{G}_{01}(\theta) &\leq T_{01} + (1 - T_{01}) = 1 \\ T_{01} + (1 - T_{01})\mathbb{G}_{00}(\theta) &\leq T_{01} + (1 - T_{01}) = 1 \\ T_{01} + T_{11} - T_{01} + (1 - T_{11})\mathbb{G}_1(\theta) = T_{11} + (1 - T_{11})\mathbb{G}_1(\theta) &\leq T_{11} + (1 - T_{11}) = 1\end{aligned}$$

It implies that $\widehat{\alpha}_A + \widehat{\zeta}^{00} \leq 1$.

Scenario B has the equilibrium if there is α such that the following holds:

$$\frac{1}{\alpha} - 1 = \frac{1 - T_{11}(1 - \mathbb{G}_1(\theta_B(\alpha))) - T_{10}\mathbb{G}_1(\theta_B(\alpha))}{T_{01}(1 - \mathbb{G}_{01}(\theta_B(\alpha))) + T_{00}\mathbb{G}_{01}(\theta_B(\alpha))} = h_B(\theta_B(\alpha))$$

Under conditions of Theorem 3, scenario B has unique equilibrium $\widehat{\alpha}_B$.

Consider case when $\mathbb{G}_{01}(\theta) \leq \mathbb{G}_{00}(\theta)$ and $G_{01}(\widehat{x}) = G_{00}(\widehat{x})$, $G_{01}(x) < G_{00}(x)$, $\forall x < \widehat{x}$, $G_{01}(x) > G_{00}(x)$, $\forall x > \widehat{x}$.

Define $f_B(\theta, \alpha) := \frac{G_{01}(\theta)}{G_1(\theta)} \frac{1-\alpha}{\alpha}$ and $f_A(\theta, \alpha, \zeta^{00}) := \frac{G_{00}(\theta)}{G_1(\theta)} \frac{\zeta^{00}}{\alpha} + \frac{G_{01}(\theta)}{G_1(\theta)} \frac{1-\alpha-\zeta^{00}}{\alpha}$. Then given α, ζ^{00} , qualification profiles under scenarios A and B are $\gamma_A(\theta) = \frac{1}{1+f_A(\theta, \alpha, \zeta^{00})}$ and $\gamma_B(\theta) = \frac{1}{1+f_B(\theta, \alpha)}$, respectively.

Because $G_{01}(\widehat{x}) = G_{00}(\widehat{x})$, $G_{01}(x) < G_{00}(x)$, $\forall x < \widehat{x}$, $G_{01}(x) > G_{00}(x)$, $\forall x > \widehat{x}$, given any α, ζ^{00} , we have $f_A(\theta, \alpha, \zeta^{00}) > f_B(\theta, \alpha)$, $\forall \theta < \widehat{x}$, $f_A(\theta, \alpha, \zeta^{00}) < f_B(\theta, \alpha)$, $\forall \theta > \widehat{x}$ and $f_A(\widehat{x}, \alpha, \zeta^{00}) = f_B(\widehat{x}, \alpha)$. By optimal equation (3), given any α, ζ^{00} , the unconstrained optimal policies $\theta_A(\alpha, \zeta^{00})$, $\theta_B(\alpha)$ under two scenarios satisfy

$$\gamma_A(\theta_A(\alpha, \zeta^{00})) = \gamma_B(\theta_B(\alpha)) = \frac{u_-}{u_+ + u_-} \iff f_A(\theta_A(\alpha, \zeta^{00}), \alpha, \zeta^{00}) = f_B(\theta_B(\alpha), \alpha) = \frac{u_+}{u_-}.$$

Denote $\beta(\widehat{x}, \alpha) := \frac{G_{01}(\widehat{x})}{G_1(\widehat{x})} \frac{1-\alpha}{\alpha} = \frac{G_{00}(\widehat{x})}{G_1(\widehat{x})} \frac{1-\alpha}{\alpha}$. It implies that $\theta_A(\alpha, \zeta^{00}) > \theta_B(\alpha)$ iff $\frac{u_+}{u_-} > \beta(\widehat{x}, \alpha)$; $\theta_A(\alpha, \zeta^{00}) < \theta_B(\alpha)$ iff $\frac{u_+}{u_-} < \beta(\widehat{x}, \alpha)$.

Since $\beta(\widehat{x}, \alpha)$ is strictly decreasing from $+\infty$ to 0 over $\alpha \in [0, 1]$, define $\bar{\alpha} := \frac{1}{1 + \frac{u_+}{u_-} \frac{G_1(\widehat{x})}{G_{01}(\widehat{x})}}$ such that $\frac{u_+}{u_-} = \beta(\widehat{x}, \bar{\alpha})$. Then $\frac{u_+}{u_-} < \beta(\widehat{x}, \alpha)$, $\forall \alpha \in [0, \bar{\alpha}]$ and $\frac{u_+}{u_-} > \beta(\widehat{x}, \alpha)$, $\forall \alpha \in (\bar{\alpha}, 1]$.

Because under Condition 1(B), we have $\frac{1}{\bar{\alpha}_B} - 1 \in (\frac{1-T_{11}}{T_{01}}, \frac{1-T_{10}}{T_{00}})$. If $\frac{u_+}{u_-} > \frac{G_{01}(\widehat{x})}{G_1(\widehat{x})} \frac{1-T_{10}}{T_{00}}$, then $\frac{u_+}{u_-} > \frac{G_{01}(\widehat{x})}{G_1(\widehat{x})} (\frac{1}{\bar{\alpha}_B} - 1) = \beta(\widehat{x}, \widehat{\alpha}_B)$ holds so that $\theta_A(\widehat{\alpha}_B, \zeta^{00}) > \theta_B(\widehat{\alpha}_B)$, $\forall \zeta^{00}$ and $\widehat{\alpha}_B \in (\bar{\alpha}, 1]$. It further implies that $\theta_A(\alpha, \zeta^{00}) > \theta_B(\alpha)$, $\forall \zeta^{00}$, $\alpha \in [\widehat{\alpha}_B, 1]$. Moreover, $(\mathbb{G}_{01}(\theta) - \mathbb{G}_{00}(\theta))(T_{01} - T_{00})\zeta^{00} < 0$ holds $\forall \zeta^{00}, \theta$. Therefore, we have

$$h_A(\theta_A(\alpha, \zeta^{00})) > h_B(\theta_B(\alpha)) \geq h_B(\theta_B(\widehat{\alpha}_B)) = \frac{1}{\widehat{\alpha}_B} - 1, \forall \alpha \in [\widehat{\alpha}_B, 1], \forall \zeta^{00}.$$

It implies that $h_A(\theta_A(\alpha, \zeta^{00}))$ has no intersection with $\frac{1}{\alpha} - 1$ over $\alpha \in [\widehat{\alpha}_B, 1]$. Therefore, $\widehat{\alpha}_A < \widehat{\alpha}_B$.

Similarly, under Condition 1(A), we have $\frac{1}{\bar{\alpha}_B} - 1 \in (\frac{1-T_{10}}{T_{00}}, \frac{1-T_{11}}{T_{01}})$. If $\frac{u_+}{u_-} > \frac{G_{01}(\widehat{x})}{G_1(\widehat{x})} \frac{1-T_{11}}{T_{01}}$, then $\theta_A(\widehat{\alpha}_B, \zeta^{00}) > \theta_B(\widehat{\alpha}_B)$, $\forall \zeta^{00}$ and $\widehat{\alpha}_B \in (\bar{\alpha}, 1]$ hold. It further implies $\theta_A(\alpha, \zeta^{00}) > \theta_B(\alpha)$, $\forall \zeta^{00}$, $\alpha \in [\widehat{\alpha}_B, 1]$. Therefore, we have

$$h_A(\theta_A(\alpha, \zeta^{00})) < h_B(\theta_B(\alpha)) \leq h_B(\theta_B(\widehat{\alpha}_B)) = \frac{1}{\widehat{\alpha}_B} - 1, \forall \alpha \in [\widehat{\alpha}_B, 1], \forall \zeta^{00}.$$

Therefore, $\widehat{\alpha}_A > \widehat{\alpha}_B$.

Using the similar analysis, the case when $\mathbb{G}_{01}(\theta) \geq \mathbb{G}_{00}(\theta)$ and $G_{01}(\widehat{x}) = G_{00}(\widehat{x})$, $G_{01}(x) > G_{00}(x)$, $\forall x < \widehat{x}$, $G_{01}(x) < G_{00}(x)$, $\forall x > \widehat{x}$ can also be proved and is omitted. \square

The Proof of Theorem 8.

Proof. In the proof, let $G_{01}(x) = G_{00}(x) = G_0(x)$ and remove subscript A to simplify the notation, i.e., $G_{1a}^A(\theta) := G_{1a}(\theta)$, $\forall a \in \{0, 1\}$.

Under Assumption 3 and the case when $G_{01}(x) = G_{00}(x) = G_0(x)$, $G_{11}(x) \neq G_{10}(x)$, use the fact that $\zeta_t^{01} + \zeta_t^{00} = 1 - \alpha_t$, $\zeta_t^{11} + \zeta_t^{10} + \zeta_t^{01} + \zeta_t^{00} = 1$, dynamics of qualification rate can be simplified

as follows, where θ_t is the threshold of the policy at t .

$$\begin{aligned} 1 - \alpha_{t+1} &= \left[(1 - T_{10})(1 - \mathbb{G}_0(\theta_t)) + (1 - T_{00})\mathbb{G}_0(\theta_t) - (1 - T_{11})(1 - \mathbb{G}_{10}(\theta_t)) \right. \\ &\quad \left. - (1 - T_{10})\mathbb{G}_{10}(\theta_t) \right] (1 - \alpha_t) + (T_{11} - T_{10})(\mathbb{G}_{11}(\theta_t) - \mathbb{G}_{10}(\theta_t))\zeta_t^{11} \\ &\quad + (1 - T_{11})(1 - \mathbb{G}_{10}(\theta_t)) + (1 - T_{10})\mathbb{G}_{10}(\theta_t) \\ \zeta_{t+1}^{11} &= \left[T_{01}(1 - \mathbb{G}_0(\theta_t)) - T_{11}(1 - \mathbb{G}_{10}(\theta_t)) \right] (1 - \alpha_t) \\ &\quad + T_{11}(1 - \mathbb{G}_{10}(\theta_t)) + T_{11}(\mathbb{G}_{10}(\theta_t) - \mathbb{G}_{11}(\theta_t))\zeta_t^{11} \end{aligned}$$

and qualification profile is reduced to $\gamma_t(x) = \frac{1}{1 + \frac{(1 - \alpha_t)\mathbb{G}_0(x)}{\mathbb{G}_{10}(x)\zeta_t^{10} + \mathbb{G}_{11}(x)\zeta_t^{11}}}$.

Scenario A has the equilibrium if there exists α, ζ^{11} such that the following holds:

$$\begin{aligned} 1 - \alpha &= \left[(1 - T_{10})(1 - \mathbb{G}_0(\theta)) + (1 - T_{00})\mathbb{G}_0(\theta) - (1 - T_{11})(1 - \mathbb{G}_{10}(\theta)) \right. \\ &\quad \left. - (1 - T_{10})\mathbb{G}_{10}(\theta) \right] (1 - \alpha) + (T_{11} - T_{10})(\mathbb{G}_{11}(\theta) - \mathbb{G}_{10}(\theta))\zeta^{11} \\ &\quad + (1 - T_{11})(1 - \mathbb{G}_{10}(\theta)) + (1 - T_{10})\mathbb{G}_{10}(\theta) \\ \zeta^{11} &= \left[T_{01}(1 - \mathbb{G}_0(\theta)) - T_{11}(1 - \mathbb{G}_{10}(\theta)) \right] (1 - \alpha) \\ &\quad + T_{11}(1 - \mathbb{G}_{10}(\theta)) + T_{11}(\mathbb{G}_{10}(\theta) - \mathbb{G}_{11}(\theta))\zeta^{11} \end{aligned}$$

Re-organize above equations,

$$\begin{aligned} \frac{1}{1 - \alpha} - 1 &= h_A(\theta_A(\zeta^{11}, \alpha)) \\ &= \frac{1 - (1 - T_{01})(1 - \mathbb{G}_0(\theta)) - (1 - T_{00})\mathbb{G}_0(\theta) - (T_{11} - T_{10})(\mathbb{G}_{11}(\theta) - \mathbb{G}_{10}(\theta))\zeta^{11}}{(1 - T_{11})(1 - \mathbb{G}_{10}(\theta)) + (1 - T_{10})\mathbb{G}_{10}(\theta) + (T_{11} - T_{10})(\mathbb{G}_{11}(\theta) - \mathbb{G}_{10}(\theta))\zeta^{11}} \\ \zeta^{11} &= \frac{[T_{01}(1 - \mathbb{G}_0(\theta)) - T_{11}(1 - \mathbb{G}_{10}(\theta))](1 - \alpha) + T_{11}(1 - \mathbb{G}_{10}(\theta))}{1 - T_{11}(\mathbb{G}_{01}(\theta) - \mathbb{G}_{11}(\theta))} = g_A(\theta_A(\zeta^{11}, \alpha)) \end{aligned}$$

where $\theta := \theta_A(\zeta^{11}, \alpha)$ for simplification. Given any ζ^{11} , $h_A(\theta(\zeta^{11}, \alpha))$ is continuous in α , there exists at least one $\bar{\alpha} := \psi(\zeta^{11})$ such that $\frac{1}{1 - \bar{\alpha}} - 1 = h_A(\theta_A(\zeta^{11}, \bar{\alpha}))$ holds.

Given any $\alpha \in [0, 1]$, $g_A(\theta_A(\zeta^{11}, \alpha)) \in (0, 1)$ is continuous in ζ^{11} , while ζ^{11} is strictly increasing from 0 to 1. Therefore, there exists at least one $\bar{\zeta}^{11} := \eta(\alpha)$ such that $\bar{\zeta}^{11} = g_A(\theta_A(\zeta^{11}, \alpha))$.

Consider a 2D plane $\{(\alpha, \zeta^{11}) : \zeta^{11} \in [0, 1], \alpha \in [0, 1]\}$, two sets of curves $\mathcal{C}_1 = \{(\alpha, \zeta^{11}) : \alpha = \psi(\zeta^{11}), \zeta^{11} \in (0, 1)\}$ and $\mathcal{C}_2 = \{(\alpha, \zeta^{11}) : \zeta^{11} = \eta(\alpha), \alpha \in (0, 1)\}$, then \mathcal{C}_1 and \mathcal{C}_2 must have at least one intersection, noted as $(\hat{\alpha}_A, \hat{\zeta}^{11})$. However, in order for $(\hat{\alpha}_A, \hat{\zeta}^{11})$ to be a valid equilibrium of the system for scenario A , it must satisfy $0 \leq 1 - \hat{\alpha}_A + \hat{\zeta}^{11} \leq 1$. We show it in the following.

Sum up above balanced equations regarding $1 - \alpha$ and ζ^{11} gives

$$\begin{aligned} &1 - \hat{\alpha}_A + \hat{\zeta}^{11} \\ &= (T_{10}\mathbb{G}_{10}(\theta) - T_{00}\mathbb{G}_0(\theta))(1 - \hat{\alpha}_A) - T_{10}(\mathbb{G}_{11}(\theta) - \mathbb{G}_{10}(\theta))\hat{\zeta}^{11} + 1 - T_{10}\mathbb{G}_{10}(\theta) \\ &= 1 - \left(\mathbb{G}_{10}(\theta)T_{10}(\hat{\alpha}_A - \hat{\zeta}^{11}) + T_{00}\mathbb{G}_0(\theta)(1 - \hat{\alpha}_A) + T_{10}\mathbb{G}_{11}(\theta)\hat{\zeta}^{11} \right) \end{aligned}$$

Since $\mathbb{G}_{10}(\theta)T_{10}(\hat{\alpha}_A - \hat{\zeta}^{11}) + T_{00}\mathbb{G}_0(\theta)(1 - \hat{\alpha}_A) + T_{10}\mathbb{G}_{11}(\theta)\hat{\zeta}^{11}$ is the convex combination of $\mathbb{G}_{10}(\theta)T_{10} \in [0, 1]$, $T_{00}\mathbb{G}_0(\theta) \in [0, 1]$ and $T_{10}\mathbb{G}_{11}(\theta) \in [0, 1]$. Therefore, $1 - \hat{\alpha}_A + \hat{\zeta}^{11} \in [0, 1]$.

Scenario B has the equilibrium if there is α such that the following holds:

$$\frac{1}{1 - \alpha} - 1 = \frac{1 - (1 - T_{01})(1 - \mathbb{G}_0(\theta_B(\alpha))) - (1 - T_{00})\mathbb{G}_0(\theta_B(\alpha))}{(1 - T_{11})(1 - \mathbb{G}_{10}(\theta_B(\alpha))) + (1 - T_{10})\mathbb{G}_{10}(\theta_B(\alpha))} = h_B(\theta_B(\alpha)).$$

Under conditions of Theorem 3, scenario B has unique equilibrium $\hat{\alpha}_B$.

Consider the case when $\mathbb{G}_{11}(x) \leq \mathbb{G}_{10}(x), \forall x$ and $G_{11}(\hat{x}) = G_{10}(\hat{x}), G_{11}(x) < G_{10}(x), \forall x < \hat{x}, G_{11}(x) > G_{10}(x), \forall x > \hat{x}$.

Define $f_B(\theta, \alpha) := \frac{G_0(\theta)}{G_{10}(\theta)} \frac{1-\alpha}{\alpha}$ and $f_A(\theta, \alpha, \zeta^{11}) := \frac{G_0(\theta)(1-\alpha)}{G_{10}(\theta)(1-\zeta^{11}-\alpha)+G_{11}(\theta)\zeta^{11}}$. Then given α, ζ^{11} , qualification profiles under scenarios A and B are $\gamma_A(\theta) = \frac{1}{1+f_A(\theta, \alpha, \zeta^{11})}$ and $\gamma_B(\theta) = \frac{1}{1+f_B(\theta, \alpha)}$, respectively. Given any α, ζ^{11} , we have $f_A(\theta, \alpha, \zeta^{11}) > f_B(\theta, \alpha), \forall \theta < \hat{x}, f_A(\theta, \alpha, \zeta^{11}) < f_B(\theta, \alpha), \forall \theta > \hat{x}$ and $f_A(\hat{x}, \alpha, \zeta^{11}) = f_B(\hat{x}, \alpha)$. By optimal equation (3), given any α, ζ^{11} , the unconstrained optimal policies $\theta_A(\alpha, \zeta^{11}), \theta_B(\alpha)$ under two scenarios satisfy

$$\gamma_A(\theta_A(\alpha, \zeta^{11})) = \gamma_B(\theta_B(\alpha)) = \frac{u_-}{u_+ + u_-} \iff f_A(\theta_A(\alpha, \zeta^{11}), \alpha, \zeta^{11}) = f_B(\theta_B(\alpha), \alpha) = \frac{u_+}{u_-}.$$

Denote $\beta(\hat{x}, \alpha) := \frac{G_0(\hat{x})}{G_{10}(\hat{x})} \frac{1-\alpha}{\alpha} = \frac{G_0(\hat{x})}{G_{11}(\hat{x})} \frac{1-\alpha}{\alpha}$. It implies that $\theta_A(\alpha, \zeta^{11}) > \theta_B(\alpha)$ iff $\frac{u_+}{u_-} > \beta(\hat{x}, \alpha)$; $\theta_A(\alpha, \zeta^{11}) < \theta_B(\alpha)$ iff $\frac{u_+}{u_-} < \beta(\hat{x}, \alpha)$.

Since $\beta(\hat{x}, \alpha)$ is strictly decreasing from $+\infty$ to 0 over $\alpha \in [0, 1]$, define $\bar{\alpha} := \frac{1}{1 + \frac{u_+}{u_-} \frac{G_{10}(\hat{x})}{G_0(\hat{x})}}$ such that $\frac{u_+}{u_-} = \beta(\hat{x}, \bar{\alpha})$. Then $\frac{u_+}{u_-} < \beta(\hat{x}, \alpha), \forall \alpha \in [0, \bar{\alpha}]$ and $\frac{u_+}{u_-} > \beta(\hat{x}, \alpha), \forall \alpha \in (\bar{\alpha}, 1]$.

Because under Condition 1(B), we have $\frac{1}{\bar{\alpha}_B} - 1 \in (\frac{1-T_{11}}{T_{01}}, \frac{1-T_{10}}{T_{00}})$. If $\frac{u_+}{u_-} < \frac{G_0(\hat{x})}{G_{10}(\hat{x})} \frac{1-T_{11}}{T_{01}}$, then $\frac{u_+}{u_-} < \frac{G_0(\hat{x})}{G_{10}(\hat{x})} (\frac{1}{\bar{\alpha}_B} - 1) = \beta(\hat{x}, \hat{\alpha}_B)$ holds so that $\theta_A(\hat{\alpha}_B, \zeta^{11}) < \theta_B(\hat{\alpha}_B), \forall \zeta^{11}$ and $\hat{\alpha}_B \in [0, \bar{\alpha}]$. It further implies that $\theta_A(\alpha, \zeta^{11}) < \theta_B(\alpha), \forall \zeta^{11}, \alpha \in [0, \hat{\alpha}_B]$. Moreover, $(T_{11} - T_{10})(\mathbb{G}_{11}(\theta) - \mathbb{G}_{10}(\theta))\zeta^{11} < 0$ holds $\forall \zeta^{11}, \theta$. Therefore, we have

$$h_A(\theta_A(\alpha, \zeta^{11})) > h_B(\theta_B(\alpha)) \geq h_B(\theta_B(\hat{\alpha}_B)) = \frac{1}{\bar{\alpha}_B} - 1, \forall \alpha \in [0, \hat{\alpha}_B], \forall \zeta^{11}.$$

It implies that $h_A(\theta_A(\alpha, \zeta^{11}))$ has no intersection with $\frac{1}{1-\alpha} - 1$ over $\alpha \in [0, \hat{\alpha}_B]$. Therefore, $\hat{\alpha}_A > \hat{\alpha}_B$.

Similarly, under Condition 1(A), we have $\frac{1}{\bar{\alpha}_B} - 1 \in (\frac{1-T_{10}}{T_{00}}, \frac{1-T_{11}}{T_{01}})$. if $\frac{u_+}{u_-} < \frac{G_0(\hat{x})}{G_{10}(\hat{x})} \frac{1-T_{10}}{T_{00}}$, then $\theta_A(\hat{\alpha}_B, \zeta^{11}) < \theta_B(\hat{\alpha}_B), \forall \zeta^{11}$ and $\hat{\alpha}_B \in [0, \bar{\alpha}]$ hold. It further implies $\theta_A(\alpha, \zeta^{11}) < \theta_B(\alpha), \forall \zeta^{11}, \alpha \in [0, \hat{\alpha}_B]$. Therefore, we have

$$h_A(\theta_A(\alpha, \zeta^{00})) < h_B(\theta_B(\alpha)) \leq h_B(\theta_B(\hat{\alpha}_B)) = \frac{1}{\bar{\alpha}_B} - 1, \forall \alpha \in [0, \hat{\alpha}_B], \forall \zeta^{11}.$$

Therefore, $\hat{\alpha}_A < \hat{\alpha}_B$.

Using the similar analysis, the case when $\mathbb{G}_{11}(x) \geq \mathbb{G}_{10}(x), \forall x$ and $G_{11}(\hat{x}) = G_{10}(\hat{x}), G_{11}(x) > G_{10}(x), \forall x < \hat{x}, G_{11}(x) < G_{10}(x), \forall x > \hat{x}$ can also be proved and is omitted.

□