The Disparate Equilibria of Algorithmic Decision Making when Individuals Invest Rationally

Lydia T. Liu
University of California, Berkeley

Ashia Wilson
Microsoft Research

Nika Haghtalab
Cornell University

Adam Tauman Kalai
Microsoft Research

Christian Borgs
Microsoft Research

Jennifer Chayes
Microsoft Research

ABSTRACT

The long-term impact of algorithmic decision making is shaped by the dynamics between the deployed decision rule and individuals' response. Focusing on settings where each individual desires a positive classification—including many important applications such as hiring and school admissions, we study a dynamic learning setting where individuals invest in a positive outcome based on their group's expected gain and the decision rule is updated to maximize institutional benefit. By characterizing the equilibria of these dynamics, we show that natural challenges to desirable long-term outcomes arise due to heterogeneity across groups and the lack of realizability. We consider two interventions, decoupling the decision rule by group and subsidizing the cost of investment. We show that decoupling achieves optimal outcomes in the realizable case but has discrepant effects that may depend on the initial conditions otherwise. In contrast, subsidizing the cost of investment is shown to create better equilibria for the disadvantaged group even in the absence of realizability.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning*; • **Applied computing** → *Law, social and behavioral sciences*; *Economics*;

KEYWORDS

fairness; machine learning; dynamics; statistical discrimination

ACM Reference Format:
Lydia T. Liu, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes. 2020. The Disparate Equilibria of Algorithmic Decision Making, when Individuals Invest Rationally. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3351095.3372861

1 INTRODUCTION

Automated decision-making systems that rely on machine learning are increasingly used for high-stakes applications, yet their long-term consequences have been controversial and poorly understood. On one hand, deployed decision making models are updated periodically to assure high performance on the target distribution. On the other hand, deployed models can reshape the underlying populations thus biasing how the model is updated in the future. This complex interplay between algorithmic decisions, individual-level responses, and exogeneous societal forces can lead to pernicious long term effects that reinforce or even exacerbate existing social injustices [13, 44]. Harmful feedback loops have been observed in automated decision making in several contexts including recommendation systems [7, 11, 38], predictive policing [18], admission decisions [5, 35], and credit markets [1, 20]. These examples underscore the need to better understand the dynamics of algorithmic decision making, in order to align decisions made about people with desirable long-term societal outcomes.

Automated decision-making algorithms rely on observable features to predict some variable of interest. In a setting such as hiring, decision making models *assess* features such as scores on standardized tests, resume, and recommendation letters, to identify individuals that are *qualified* for the job. However, equally qualified people from different demographic groups tend to have different features, due to implicit societal biases (e.g., letter writers describe competent men and women differently), gaps in resources (e.g., affluent students can afford different extra-curriculars) and even distinct tendencies in self-description (e.g., gender can be inferred from biographies [16]). Therefore, a model's ability to identify qualified individuals can widely vary across different groups.

The deployed model's ability to identify qualified members of a group affects an individual's incentive to *invest* in their qualification. This is because one's decision to acquire qualification—not observed directly by the algorithm—comes at a cost. Moreover, individuals that are identified by the model as qualified (whether or not they are truly qualified) receive a reward. Consequently, people invest in acquiring qualifications only when their expected reward from the assessment model beats the investment cost.

Rational individuals are aware that upon investing they would develop features that are similar to those of qualified individuals in their group, so they gauge their own expected reward from investing by the observed rewards of their group.[1] If qualified people from one group are not duly identified and rewarded, fewer people from that group are incentivized to invest in qualifications in the future. This reduces the overall fraction of qualified people in that group, or the *qualification rate.* As the assessment model is updated to maximize overall institutional profit on the new population distribution, it may perform even more poorly on qualified individuals

[1] Strong group identification effects can also be seen in empirical studies [24].

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02.
https://doi.org/10.1145/3351095.3372861

from a group with relatively low qualification rate, further reducing the group's incentive to invest.

To understand and mitigate the challenges to long-term welfare and fairness posed by such dynamics, we propose a formal model of sequential learning and decision-making where at each round a new batch of individuals rationally decide whether to invest in acquiring qualification and the institution updates its assessment rule (a classifier) for assessing and thus rewarding individuals. We study the long-term behavior of these dynamics by characterizing their equilibria and comparing these equilibria based on several metrics of social desirability. Our model can be seen as an extension of Coate and Loury [10]'s widely cited work to explicitly address heterogeneity in observed features across groups. While Coate and Loury [10]'s model focuses on a single-dimensional feature space, i.e., scores, and assessment rules that act as thresholds on the score, our model considers general, possibly high-dimensional, feature spaces and arbitrary assessment rules, which are typical in high-stakes domains such as hiring and admissions.

We find that two major obstacles to obtaining desirable long-term outcomes are heterogeneity across groups and lack of realizability within a group. *Realizability*—the existence of a (near) perfect way to assess qualifications of individuals from visible features—leads to equilibria that are (near) optimal on several metrics, such as the resulting qualification rates, their uniformity across groups, and the institution's utility. We study (near) realizability and the lack thereof in Sections 3 and 5 respectively. *Heterogeneity across groups*, i.e., lack of a single assessment rule that perfectly assesses individuals from all groups, necessitates tradeoffs in the quality of equilibria across different groups. We study heterogeneity, as well as interventions for mitigating its negative effects, in Section 4. In Section 6, we empirically study a more challenging setting where the groups are heterogeneous as well as highly non-realizable, via simulations with a FICO credit score dataset [42] that has been widely used for illustration in the algorithmic fairness literature.

Interventions. To mitigate the aforementioned tradeoffs, we consider two common interventions: decoupling the decision policy by group and subsidizing the cost of investment, especially when the cost distribution inherently differs by group. Our model of dynamics sheds a different light on these interventions, complementary to previous work. We show that decoupling [17]—using group-specific assessment rules—achieves optimal outcomes when the problem is realizable within each group, but can significantly hurt certain groups when the problem is non-realizable and there exist multiple equilibria after decoupling. In particular, decoupling can hurt a group with low initial qualification rate if the utility-maximizing assessment rule for a single group is more disincentivizing to individuals than a joint assessment rule, thereby reinforcing the status quo and preventing the group from reaching an equilibrium with higher qualification rate.

We also study subsidizing individuals' investment cost (e.g. subsidizing tuition for a top high school), especially when the cost distribution is varied across different groups. We find that these subsidies increase the qualification rate of the disadvantaged group at equilibrium, regardless of realizability. We note that our subsidies, which affect the qualification of individuals directly, are different than those studied under strategic manipulation [26] that involve subsidizing individual's cost to manipulate their features *without changing the underlying qualification* (e.g. subsidizing SAT exam preparation without changing the student's qualification for college) and could have adverse effects on disadvantaged groups. Instead, our theoretical findings resonates with extensive empirical work in economics on the effectiveness of subsidizing opportunities for a disadvantaged group to directly improve their outcomes, such as moving to better neighborhoods to access better educational and environmental resources [8].

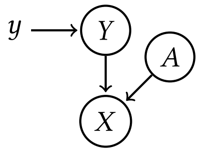

Figure 1: Causal graph for the individual investment model. The individual intervenes on the node for qualification, Y—this corresponds to do$(Y = y)$)—which then affects the distribution of their features X, depending on the group A.

Related work. Our work is related to a rich body of work on algorithmic fairness in dynamic settings [23, 25, 33, 37, 45], strategic classification [26, 31, 36], as well as statistical discrimination in economics [2, 3, 10]. We present a detailed discussion of the similarities and differences in Section 7.

2 A DYNAMIC MODEL OF ALGORITHMIC DECISION MAKING

In this section we introduce a model of automated decision making with feedback. We first introduce the notation used throughout the paper and then describe the details of the interactions between individuals and an institution, and the resulting dynamical system.

2.1 Notation

We consider an instance space $\mathcal{X}$, where $X \in \mathcal{X}$ denotes the features of an individual that are observable by the institution. We also consider a label space $\mathcal{Y} = \{0, 1\}$ where $Y = 1$ indicates that an individual has the qualifications desired by the institution and $Y = 0$ otherwise. We denote the set of all protected/group attributes by $\mathcal{A}$ where $A \in \mathcal{A}$ denotes an individual's protected attribute. We denote the group proportions by $n_a := \mathbb{P}(A = a)$ for all $a \in \mathcal{A}$. Furthermore, we denote the qualification rate in group $a \in \mathcal{A}$ by $\pi_a := \mathbb{P}(Y = 1 \mid A = a)$. An individual from group $A = a$ who has acquired label $Y = y$ (to become qualified or not)[2] receives features X distributed according to $\mathbb{P}(X = x \mid Y = y, A = a)$. This is illustrated in Figure 1.

We also consider a set of parameters Θ that are used for assessing qualifications. We use $\hat{Y}_\theta \in \mathcal{Y}$ parameterized by $\theta \in \Theta$ to denote the *assessed qualification* of an individual. We assume that $\hat{Y}_\theta$ only depends on the features X, which may or may not contain A or its proxies. In later sections, we will also discuss interventions that allow us to use $\hat{Y}_\theta$ that explicitly depends on group membership A. We respectively define the *true positive rate* and *false positive rate*

[2] This can be seen as the individual performing a do-intervention on Y [see e.g., 39]. Thus we may write do$(Y = 1)$ for making the decision to acquire qualifications. Our model (Figure 1) assumes that Y is not the child of any node, so we have $\mathbb{P}(\cdot \mid \text{do}(Y = y)) = \mathbb{P}(\cdot \mid Y = y)$. Hence we drop the do-operator whenever we condition on Y.

of $\theta \in \Theta$ on group $a \in \mathcal{A}$ by

$$\mathrm{TPR}_a(\theta) = \mathbb{P}(\hat{Y}_\theta = 1 \mid Y = 1, A = a), \text{ and}$$
$$\mathrm{FPR}_a(\theta) = \mathbb{P}(\hat{Y}_\theta = 1 \mid Y = 0, A = a).$$

2.2 Model Description

Individual's Rational Response. We consider a setting where an individual decides whether to acquire qualifications, that is, to invest in obtaining label $Y = 1$, prior to observing their feature X. The decision to acquire qualification depends on the qualification assessment rule $\theta \in \Theta$ currently implemented by the institution. We will characterize the groups' qualification rates as the *best-response* to θ by function $\pi^{br}(\theta) = (\pi_a^{br}(\theta))_{a\in\mathcal{A}}$.

To get label $Y = 1$ an individual has to pay a cost $C > 0$. In any group, C is distributed randomly according to the cumulative distribution function (CDF), $G(\cdot)$.[3] After deciding whether to acquire qualifications, an individual gets features X and is assessed by θ. An individual (from any group and regardless of actual qualification) receives a payoff of $w > 0$ if they are assessed to be qualified and payoff of 0 otherwise. Therefore, the expected utility an individual from group a receives from acquiring qualification $Y = 1$ is $w\mathbb{P}[\hat{Y}_\theta = 1 | Y = 1, A = a] - C = w\mathrm{TPR}_a(\theta) - C$ whereas the expected utility for not acquiring the qualification is $w\mathbb{P}[\hat{Y}_\theta = 1 | Y = 0, A = a] = w\mathrm{FPR}_a(\theta)$. Given the qualification assessment parameter $\theta \in \Theta$, an individual from group a acquires qualification if and only if the benefit outweighs the costs, that is

$$w(\mathrm{TPR}_a(\theta) - \mathrm{FPR}_a(\theta)) > C. \tag{1}$$

Then each group's qualification rate as a function of a qualification assessment parameter θ is

$$\begin{aligned}\pi_a^{br}(\theta) := \mathbb{P}(Y = 1 \mid A = a) &= \mathbb{P}(C < w(\mathrm{TPR}_a(\theta) - \mathrm{FPR}_a(\theta)))\\ &= G(w(\mathrm{TPR}_a(\theta) - \mathrm{FPR}_a(\theta))).\end{aligned}$$

Institution's Rational Response We consider an institution that has to choose a qualification assessment parameter for accepting individuals to maximize its utility. We assume that the institution gains $p_{\mathrm{TP}} > 0$ for accepting a qualified individual and loses $c_{\mathrm{FP}} > 0$ for accepting an unqualified individual. Then the expected utility of the institution for applying parameter θ is

$$\begin{aligned}&p_{\mathrm{TP}}\mathbb{P}(\hat{Y}_\theta = 1, Y = 1) - c_{\mathrm{FP}}\mathbb{P}(\hat{Y}_\theta = 1, Y = 0)\\ &= p_{\mathrm{TP}} \sum_{a\in\mathcal{A}} \mathrm{TPR}_a(\theta)\pi_a n_a - \sum_{a\in\mathcal{A}} c_{\mathrm{FP}}\mathrm{FPR}_a(\theta)(1 - \pi_a)n_a.\end{aligned}$$

This illustrates that the utility maximizing parameter is a function of $\pi = (\pi_a)_{a\in\mathcal{A}}$, i.e., the rate of qualification in each group. We denote this function by $\theta^{br}(\pi)$, defined as follows:

$$\theta^{br}(\pi) := \operatorname*{argmax}_{\theta\in\Theta}\, p_{\mathrm{TP}} \sum_{a\in\mathcal{A}} \mathrm{TPR}_a(\theta)\pi_a n_a - \sum_{a\in\mathcal{A}} c_{\mathrm{FP}}\mathrm{FPR}_a(\theta)(1-\pi_a)n_a.$$

To ensure the above object (and the resulting dynamics) are well-defined, when multiple parameters θ achieve the optimal utility we assume that $\theta^{br}(\pi)$ is uniquely defined using a fixed and well-defined tie-breaking function.

Throughout this paper we assume that the institution has exact knowledge of many quantities such as $\mathrm{TPR}_a(\theta)$, $\mathrm{FPR}_a(\theta)$, and n_a. In a nutshell, we assume that we have infinitely many samples from the underlying distributions. We discuss this further in Section 8, and leave the finite sample version of these results to future work.

Although we choose not to focus on game-theoretical aspects in this work, we note that our model can be thought of as a large game [27] or a game with a continuum of players [41].

Dynamical System and Equilibria. We are primarily interested in the evolution of qualification rate, π, over time. Given a current rate of qualification π the assessment parameter used by the institution in the next step is $\theta^{br}(\pi)$, which in turn leads to a qualification rate of $\pi^{br}(\theta^{br}(\pi))$ in the next step. Therefore, we define a dynamical system for a given initial state $\pi(0)$ such that at time t we are in state $\pi(t) = \Phi(\pi(t-1))$, where $\Phi = \pi^{br} \circ \theta^{br}$.

We say that the aforementioned dynamical system is at equilibrium if $\pi = \Phi(\pi)$. Equivalently, we are at an equilibrium if $\pi = \lim_{n\to\infty} \Phi^n(\pi(0))$ is well-defined for some $\pi(0)$, where Φ^n is an n-fold composition of Φ. We call such values of π *equilibria*, or equivalently, *fixed points* of Φ.

In general, Φ may have multiple fixed points that demonstrate different characteristics. We therefore compare the fixed points of Φ on several metrics of societal importance.

(1) *Stability:* We say that an equilibrium π^* is stable if there is a non-zero measure set of initial states $\pi(0) \in [0, 1]$ for which $\pi^* = \lim_{n\to\infty} \Phi^n(\pi(0))$. In particular, if there exists a neighborhood around π^* such that all points converge to π^* under the dynamics, we say that π^* is *locally stable*. As such, stable fixed points are robust to small perturbations in the qualification rate, which can occur due to random measurement errors.
(2) *Qualification Rate of Group a:* Recall that the qualification rate, π_a, is the fraction of individuals in group a who invested in qualifications. Since it is more desirable to have a high qualification rate in each group, we may compare equilibria based on π_a. We refer to $G(w)$ as the optimal qualification rate in group a, which is the maximum achievable qualification rate corresponding to the perfect assessment rule.[4]
(3) *Balance:* We may be interested in equilibria where the qualification rate is similar across groups, that is, to prioritize equilibria with smaller $\max_{a_1,a_2\in\mathcal{A}} |\pi^*_{a_1} - \pi^*_{a_2}|$. When this quantity is 0 we say that π^* is *fully balanced.*
(4) *Institutional utility:* We may compare equilibria based on their corresponding institution utility.

2.3 Examples From the Real World

Let us instantiate our model in the setting of two important applications from the real world.

College Admissions. Consider the college admission setting, where X corresponds to the features that the college can observe, e.g., a candidate's test scores and letters of recommendation. Y indicates whether the candidate meets the qualifications required to succeed in the program. C is the cost of investing in the qualifications, e.g., the money and opportunity cost of studying or taking additional courses to obtain the required qualifications. A

[3] For the rest of this work, unless otherwise stated, we assume that the distribution of costs, G, is the same for every group. In Section 4.3 and 6, we will consider the implications of having different cost distributions by group.

[4] If group a has a group-specific cost distribution, G_a, then we refer to $G_a(w)$ as the optimal qualification rate in group a.

candidate from group a will develop features from distribution $\mathbb{P}[X = x|Y = y, A = a]$, where $y = 1$ indicates a qualified candidate. The differences in the feature distribution between groups can be attributed to several factors such as resources that are available to different groups, e.g., letters of recommendations for qualified female and male candidates often emphasize different traits. θ is the decision parameter used by the college, e.g., $\hat{Y}_\theta = 1$ when the candidate has SAT score of > 1400 and an excellent recommendation letter. The college accepts applicants by trading off between the utility gain, p_{TP}, of admitting qualified candidates and utility cost, c_{FP}, of admitting an unqualified candidates. The candidate is incentivized to acquire the qualifications for the college based on the long term benefit (described in Equation (1)) that depends on their expected gain w from completing a college degree and how likely it is to be admitted to college for a qualified or unqualified member of the group the candidate belongs to.

Hiring. Consider the hiring setting, where X corresponds to the features that the firm can observe, e.g., a candidate's CV. Y indicates whether the applicant meets the qualifications required by the firm, e.g., having the required knowledge and the ability to work in a team. C is the cost of acquiring the qualifications required by the firm, e.g., the (monetary and opportunity) cost of acquiring a college degree or working on a team project. Parameter θ is the hiring parameter used by the firm, e.g., $\hat{Y}_\theta = 1$ when the applicant has a software engineering degree and two years of experience. The firm accepts candidates according to utility maximization involving p_{TP}, the profit from hiring a qualified candidate, and c_{FP}, the cost of hiring an unqualified candidate e.g., the loss in productivity or the the cost to replace the employee. The candidate is incentivized to acquire the qualifications for the job based on factors including their expected salary w and how likely it is to be hired by the firm given how the firm has hired qualified or unqualified candidates from the group the candidate belongs to (Eq. (1)).

We also consider a stylized example of lending in Section 6.

3 IMPORTANCE OF (NEAR) REALIZABILITY

We start our theoretical investigation of dynamic algorithmic decision making with the classical model of realizability. In the theory of machine learning, a distribution is called realizable if there is a decision rule in the set Θ whose error on the distribution is 0. Analogously, we call a setting *realizable* when there is a decision rule $\theta^{\text{opt}} \in \Theta$ that perfectly classifies every individual from every group, that is $\text{TPR}_a(\theta^{\text{opt}}) = 1$ and $\text{FPR}_a(\theta^{\text{opt}}) = 0$ for all $a \in \mathcal{A}$. Realizability is a widely used assumption and is the basis of seminal works such as Boosting [19]. At a high level, realizability corresponds to the assumption that there is an unknown ground truth assessment rule, for example, in a hypothetical setting where x includes all the information that is sufficient for assessing one's qualification, and the chosen set of decision rules is rich enough to contain it.

In static realizable applications of machine learning, the goal is to (approximately) recover θ^{opt} from data. We show that in the our dynamic setting, under realizability, the unique non-zero equilibrium of Φ is where individuals respond to θ^{opt}. Furthermore, each group attains their optimal qualification rate at this equilibrium.

Proposition 3.1 (Perfect classfication). *If there exists $\theta \in \Theta$ such that $\text{TPR}_a(\theta) = 1$ and $\text{FPR}_a(\theta) = 0$ for all $a \in \mathcal{A}$, then there is a unique non-zero equilibrium with $\pi^*_a = G(w)$ for all $a \in \mathcal{A}$.*

While realizability is a common assumption in the theory of machine learning, it rarely captures the subtleties that exist in automated decision making in practice. Next, we consider a mild relaxation of realizability and consider a setting where a near-perfect decision rule $\theta \in \Theta$ exists such that $\text{TPR}_a(\theta) \geq 1 - \epsilon$ and $\text{FPR}_a(\theta) \leq \epsilon$. As we show (and prove in Appendix A), when there is a single near-realizable group the main message of Proposition 3.1 remains effectively the same. That is, all equilibria that are reachable from initial points that are not too extreme approximately maximize the group's qualification rate.

Theorem 3.2 (Equilibria under near-realizability). *Let $|\mathcal{A}| = 1$ and assume that $p_{\text{TP}} = c_{\text{FP}} = 1$. Assume that for fixed $\epsilon \in (0, 1)$, $s \in (0, 1/2)$, G is L_G-Lipschitz with property that $1 - s \geq G(w) \geq s + \frac{L_G w\epsilon}{s}$, and there is $\theta \in \Theta$ such that*

$$\text{TPR}(\theta) \geq 1 - \epsilon \text{ and } \text{FPR}(\theta) \leq \epsilon.$$

Then for any initial investment $\pi(0) \in [s, 1-s]$, $\pi^* = \lim_{n\to\infty} \Phi^n(\pi(0))$ is such that

$$\pi^* \geq G(w(1 - \epsilon/s)).$$

A nice aspect of the above results is that the assumption of realizability or near-realizability can be validated from the data. That is, the decision maker can compute whether there is $\theta \in \Theta$ such that $\text{TPR}(\theta) \geq 1 - \epsilon$ and $\text{FPR}(\theta) \leq \epsilon$. If so, then the decision maker can rest assured that the dynamical system is on the path towards achieving near optimal investment by the individuals. Another nice aspect of these results is the characterization of the equilibria in terms of the CDF of the cost distribution. This allows us to use this framework for studying interventions that change the cost function directly. One such intervention is subsidizing the cost for individuals so that the cumulative distribution function of the cost, given by $G(x)$, is increased by a sufficient amount at every cost level x. The following corollary, proved in Appendix B, shows that under this kind of subsidy, the equilibria reached by the dynamics will have higher qualification rate than any fixed point of the dynamics before subsidy, as long as the initial points are not too extreme. As we are considering different cost distribution functions in the following corollary, we denote the dynamics corresponding to cost distribution function G as Φ_G.

Corollary 3.3 (Subsidizing the cost of investment). *Let $|\mathcal{A}| = 1$ and assume that $p_{\text{TP}} = c_{\text{FP}} = 1$. Assume that for fixed $\epsilon \in (0, 1)$, $s \in (0, 1/2)$, G is L_G-Lipschitz with property that $1 - s \geq G(w) \geq s + \frac{L_G w\epsilon}{s}$, and there is $\theta \in \Theta$ such that*

$$\text{TPR}(\theta) \geq 1 - \epsilon \text{ and } \text{FPR}(\theta) \leq \epsilon.$$

Let $\pi^* > 0$ be a fixed point of the dynamics Φ_G. Suppose $\bar{G}$ is a strictly increasing, $L_{\bar{G}}$-Lipschitz CDF such that $1 - s \geq \bar{G}(x) \geq s + \frac{L_{\bar{G}} w\epsilon}{s}$ and $\bar{G}(x(1 - \epsilon/s)) \geq G(x)$ for all x in the domain of G. Then for any initial investment $\pi(0) \in [s, 1 - s]$, there exists a $\bar{\pi} \geq \pi^*$, such that $\bar{\pi} = \lim_{n\to\infty} \Phi^n_{\bar{G}}(\pi(0))$.

4 GROUP REALIZABILITY

In this section, we investigate how the nature of equilibria evolves as the assumption of realizability is relaxed to allow for heterogeneity across groups. Specifically, we consider the case where there exists a perfect assessment rule for each group, but not when the groups are combined. We call this “group-realizability”. Our results illustrate that without realizability or near-realizability, the utility-maximizing assessment rule can be very sensitive to the relative qualification rates in different groups, resulting in *multiple* equilibria, at which groups may experience disparate outcomes.

In sections 4.1 and 4.2, we study group-realizability under two different and complementary settings. The first setting considers features that are drawn from a multivariate Gaussian distribution and assumes that in each group the qualified individuals are perfectly separated from unqualified ones by a group-specific hyperplane. This is a benign setting where no group is inherently disadvantaged — group features and performance of assessment rules are symmetric up to a reparameterization of the space. The second setting considers features that are uniformly distributed scalar scores and assumes that qualified and unqualified individuals in a group are separated by a group-specific threshold, where one is higher than the other. - This model captures the natural setting where the feature (score) and assessment rules inherently favor one group, e.g., SAT scores are known to be skewed by race [6]. We use the aforementioned stylized settings to demonstrate the salient characteristics of equilibria that one might anticipate under group-realizability. We find that stable equilibria tend to favor one group or the other. This is especially surprising in the multivariate Gaussian case where the two groups are identical up to a change in the representation of the space. We also study the existence of balanced equilibria, where both groups acquire qualification at the same rate. We find that when balanced equilibria exist they tend to be unstable, that is, no initial qualification rate (except for the balanced equilibrium itself) will converge to the balanced equilibrium under the dynamics.

We consider two natural interventions in overcoming the challenges of group-realizability as outlined above. As group-realizability poses even greater challenges when the costs of investment are unequally distributed *between* groups, in Section 4.3 we consider the impact of subsidizing the cost of acquiring qualification for one group. In Section 4.4, we consider the impact of decoupling, that is, we allow the institution to use different assessment rules for different groups assuming the group attributes are available. This is in contrast to the typical setting where institutions are constrained to using the same assessment rule across all groups, which may be the case when data on the protected attribute is not available or when the use of protected attributes for assessment is regulated.

4.1 Uniformly Distributed Scalar X

We consider $\mathcal{X} = [0, 1]$, the class of assessment paramters $\Theta = [0, 1]$, and assessment decision $\hat{Y}_h = \mathbf{1}\{X > h\}$ for all $h \in \Theta$ that represent all threshold decision policies. Consider two groups a_1, a_2. Let X be a score that is uniformly distributed over $[0, 1]$ where in group a_i those with score more than h_i are qualified and those with score at most h_i are unqualified. This is depicted in Figure 2 (right). Formally,

$$\mathbb{P}(X = x \mid Y = y, A = a_i) = \begin{cases} \mathbf{1}\{x > h_i\}/(1 - h_i) & \text{for } y = 1 \text{ and} \\ \mathbf{1}\{x \le h_i\}/h_i & \text{for } y = 0 \end{cases}.$$

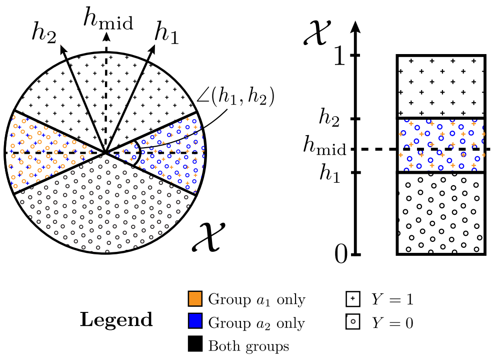

Figure 2: Equilibria in the Multivariate Gaussian case (left) and the Uniform case (right)

We make the following assumption to simplify notation.

Assumption 1. *We assume $n_{a_1} \cdot p_{\text{TP}} = n_{a_2} \cdot c_{\text{FP}}$. We also assume that the cost for acquiring qualifications is uniformly distributed on $[0, 1]$ (i.e. $G(c) = c$) in both groups.*[5]

We show that when w is in a certain range, there are two *unbalanced stable* equilibria corresponding to assessment parameters h_1 or h_2, which respectively lead to the optimal qualification rate for groups a_1 or a_2 but low qualification rate for the other group. There is also a more *balanced* but *unstable* equilibrium at some threshold h_{mid} between h_1 and h_2. Outside of this range of w, there is only one equilibrium in which one of the groups achieves its optimal qualification rate. These findings are summarized in the following two propositions.

Proposition 4.1. *Define* $g := \frac{(1-h_1)(-wh_2^2+h_2(1-h_1)-wh_1(1-h_1))}{w((1-h_1)^2-h_2^2))}$. *Note that $g \in (0, h_2 - h_1)$ for any w. Let $w \in (w_l, w_u)$ where*

$$w_l := \frac{(1 - h_1)^2}{(1 - h_2)h_2 + (1 - h_1)^2}, \; w_u := \frac{h_2(1 - h_1)}{h_2^2 + h_1(1 - h_1)}. \tag{2}$$

Given Assumption 1, there exists two stable equilibria at

$$h = h_1, \qquad \pi_{a_1} = w, \qquad \pi_{a_2} = w \cdot \frac{h_1}{h_2}, \text{ and} \tag{3}$$

$$h = h_2, \qquad \pi_{a_1} = w \cdot \frac{1 - h_2}{1 - h_1}, \qquad \pi_{a_2} = w, . \tag{4}$$

and a unique non-zero unstable equilibrium at

$$h = h_{\text{mid}} := h_1 + g, \quad \pi_{a_1} = w \cdot \frac{1 - h_1 - g}{1 - h_1}, \quad \pi_{a_2} = w \cdot \frac{h_1 + g}{h_2}.$$

When $w = 1 - h_1$, the unstable equilibrium is fully balanced.

Proposition 4.2. *Given Assumption 1 when $w < w_l$ there exists one stable equilibrium defined by Equation 4, and when $w > w_u$ there exists one stable equilibrium defined by Equation 3.*

The details of the proofs are presented in Appendix C. At a high level, if the wage is not too low or too high, both thresholds h_1 and h_2 correspond to stable equilibria, at which either group a_1 or a_2 is perfectly classified. The equilibrium corresponding to h_{mid}, where the classifier has the same true positive and false positive rates in both groups, is unstable and subsequently harder to achieve.

[5] Our results also generalize to the setting where the CDF for the cost $G : [0, 1] \to [0, 1]$ is an arbitrary strictly increasing function.

In Table 1, we compare these equilibria in terms of metrics introduced in Section 2, under the assumptions of Proposition 4.1. We use standard notation $\succ$ and $\sim$ to denote preference and indifference respectively. For example, we find that in terms of balance in qualification rates, the stable equilibrium associated with h_1 is more balanced that the stable equilibrium associated with h_2, but both are always less balanced unstable equilibrium associated with h_{mid}. Details of the computation are deferred to Table C.1 in Appendix C.

	Ranking of Equilibria
Stability	h_1, h_2 are stable. h_{mid} is unstable
Qualification rate of group a_1	$h_1 \succ h_{\text{mid}} \succ h_2$
Qualification rate of group a_2	$h_2 \succ h_{\text{mid}} \succ h_1$
Balance of qualification rates	$h_{\text{mid}} \succ h_1 \succ h_2$
Institution's Utility	no ranking

Table 1: Comparison of equilibria for uniform features. In this table we refer to each equlibria using the associated threshold decision policy.

4.2 Multivariate Gaussian X

We consider $\mathcal{X} = \mathbb{R}^d$ and $\Theta = S_{d-1}$, where S_{d-1} is the set of d-dimensional unit vectors. Let $\hat{Y}_h = \mathbf{1}\{X^\top h \geq 0\}$ for all $h \in \Theta$ denote separating hyperplane policies and $\angle_{h,h'} := \frac{1}{\pi} \arccos(\frac{h^\top h'}{\|h\|\|h'\|})$ denote the angle between two vectors, normalized by the constant π. We consider two groups a_1 and a_2 associated respectively with vectors h_1 and h_2, such that $\angle_{h_1,h_2} \neq 0$. We assume the groups have equal size, i.e., $n_{a_1} = n_{a_2}$. For each group, the feature distribution is a d-dimensional spherical Gaussian centered at the origin such that the qualified individuals are in halfspace $\mathbf{1}\{X^\top h_i \geq 0\}$ and the unqualified individuals in halfspace $\mathbf{1}\{X^\top h_i < 0\}$. Formally, for $x \in \mathbb{R}^d$ and $i \in \{1, 2\}$,

$$\mathbb{P}(X = x \mid Y = y, A = a_i) = \begin{cases} 2\phi(x)\mathbf{1}\{x^\top h_i \geq 0\} & \text{for } y = 1 \text{ and} \\ 2\phi(x)\mathbf{1}\{x^\top h_i < 0\} & \text{for } y = 0, \end{cases}$$

where $\phi(x)$ is the density of the spherical d-dimensional Gaussian. This is depicted in Figure 2 (left).

Assumption 2. *We assume that the CDF for the cost of acquiring qualifications is a strictly increasing function $G : [0, 1] \to [0, 1]$ and is the same in both groups.*

As we will see, the relative gain (loss) of the institution for accepting a qualified (unqualified) individual, that is $p_{\text{TP}}/c_{\text{FP}}$, plays a role in the nature of the equilibria. The following proposition characterizes the equilibria when this value is strictly positive, that is, when the benefit of true positives outweighs the cost of false positives. Notably, similar to the previous setting of uniform scores, the current setting also has two stable equilibria that each favor one group at the expense of the other, as well as a balanced equilibrium that is unstable.

Proposition 4.3. *Given Assumption 2 and $p_{\text{TP}} > c_{\text{FP}}$, there exists two stable equilibria, at*

$$h = h_1, \quad \pi_{a_1} = G(w) \qquad \pi_{a_2} = G\left(w \cdot (1 - 2\angle_{h_1,h_2})\right),$$
$$h = h_2, \quad \pi_{a_1} = G\left(w(1 - 2\angle_{h_1,h_2})\right) \quad \pi_{a_2} = G(w).$$

There is a unique non-zero unstable equilibrium at

$$h = h_{\text{mid}}, \quad \pi_{a_1} = G\left(w(1 - \angle_{h_1,h_2})\right) \quad \pi_{a_2} = G\left(w(1 - \angle_{h_1,h_2})\right),$$

where $h_{\text{mid}} := \frac{h_1+h_2}{\|h_1+h_2\|}$.

Let us briefly comment on the high level proof idea and defer the full argument to Appendix D. Since $p_{\text{TP}} > c_{\text{FP}}$, the institution cares more about accepting true positives than avoiding false positives. Therefore, the utility-maximizing h is determined by the group that has a higher qualification rate and thus has a higher fraction of positives — this is h_1 (resp. h_2) whenever $\pi_{a_1} > \pi_{a_2}$ (resp. $\pi_{a_1} < \pi_{a_2}$). When qualification rates are equal between the two groups, the institution maximizes its utility at any h that is a convex combination of h_1 and h_2, but the unique h that would induce equal qualification rate is $h = h_{\text{mid}}$, where the classifier has the same true positive and false positive rates in both groups.

An unfortunate implication of this result is that the dynamics will always converge to an unbalanced qualification rate, except when the initial levels of investment are exactly the same. Even though a fully balanced equilibrium exists, it is unstable and therefore not robust to small perturbations in either the qualification rates or the classifier, which in practice is unavoidable given sampling noise.

In Table 2, we compare these equilibria in terms of metrics introduced in Section 2. For example, we find that in terms of institutional utility, the stable equilibria associated with h_1 and h_2 are equally preferred, and are both strictly preferred to the unstable equilibrium associated with h_{mid}. This implies that the institution has no incentive at all to keep the dynamics at the unstable equilibrium, even though it induces balanced investment. Exact values are deferred to Table D.1 in Appendix D.

	Ranking of Equilibria
Stability	h_1, h_2 are stable. h_{mid} is unstable
Qualification rate of group a_1	$h_1 \succ h_{\text{mid}} \succ h_2$
Qualification rate of group a_2	$h_2 \succ h_{\text{mid}} \succ h_1$
Balance of qualification rate	$h_{\text{mid}} \succ h_1 \sim h_2$
Institution's Utility	$h_1 \sim h_2 \succ h_{\text{mid}}$

Table 2: Comparison of equilibria for Multivariate Gaussian features. In this table we refer to each equlibria using the associated hyperplane.

Interestingly, when $p_{\text{TP}} < c_{\text{FP}}$, there are no stable equilibria; instead there exists a stable limit cycle between h_1 and h_2. This is stated informally in the following proposition.

Proposition 4.4. *Given Assumption 2 and $p_{\text{TP}} < c_{\text{FP}}$, there exists no stable equilibria. Instead there exists a limit cycle and one non-trivial unstable equilibrium.*

Intuitively, the cycle is caused by misaligned incentives between the institution and the individuals. Since the institution finds false positives more costly than false negatives, it prefers the hyperplane that classifies more false positives correctly. At each time step, it will choose the hyperplane associated with the group that has a lower qualification rate, prompting that group to invest more in the next time step. Strikingly, even a simple group-realizable model involving multivariate Gaussian distributions demonstrates a large range of limiting behaviors. In Section 6, we also observe the existence of limit cycles in simulations with real data distributions.

4.3 Different Costs of Investment by Group

Thus far we have assumed that all groups have the same distribution of the cost of investment, G. In reality, the cost of investment may be distributed differently in each group; a disadvantaged group might on average experience higher (monetary or opportunity) costs. For example, low income families who may have to take out loans to pay for college tuition incur high interest rates. This is a compelling setting that reflects deep-seated disparities in access to opportunity between demographic groups in the real world; an analogous setting has been considered by works on strategic classification, where the costs for manipulating features is posited to differ across groups [26, 36].

In this section, we consider the ramifications of differences in investment cost across groups, focusing on the setting of Section 4.2. We show that the disadvantage from having higher costs is amplified under group-realizability. Specifically, suppose that group a_1 (resp. a_2) has costs distributed according to cumulative distribution function G_1 (resp. G_2), and that group a_1 is disadvantaged in terms of costs. The following result observes that if G_1 sufficiently dominates G_2, then there exists no stable equilibrium that encourages optimal investment from group a_1 and no equilibrium that is balanced for both groups, in sharp contrast to the characterization in Proposition 4.3. The proof is deferred to Appendix E.

Proposition 4.5. *Consider the multi-variate Gaussian setting of Section 4.2. Suppose G_1 and G_2 are such that $G_1(w) < G_2(w(1 - 2\angle_{h_1,h_2}))$, then there exists a single non-trivial equilibrium at h_2, which is also stable. The level of investment by group a_1 (resp. a_2) is $G_1(w(1 - 2\angle_{h_1,h_2})$ (resp. $G_2(w)$).*

Effect of subsidies. In this situation, an intervention that would effectively raise the equilibrium level of investment by the disadvantaged group is to subsidize the cost of investment. In particular, as long as we replace G_1 with a stochastically dominated distribution $\bar{G}_1$ such that $\bar{G}_1 > G_2(w(1 - 2\angle_{h_1,h_2}))$, under the new dynamics Φ^{sub}, h_1 will again be a stable equilibrium, and there will also exist a more balanced, unstable equilibrium at $h = \bar{h}_{\text{mid}}$, which is some convex combination of h_1 and h_2. At all equilibria of Φ^{sub}, group a_1 will have higher levels of investment than $G_1(w(1 - 2\angle_{h_1,h_2}))$.

However, this improvement may come at a cost to the advantaged group, since Φ^{sub} has multiple equilibria and some of them have group a_2 investing less than $G_2(w)$. Still one might argue that the equilibria of Φ^{sub} are more equitable, since the dynamics without subsidies always result in optimal investment by group a_2 and low investment by group a_1.

4.4 Decoupling the Assessment Rule by Group

The models we studied in Sections 4.2 and 4.1 suggest that applying the same, or "joint", assessment rule to heterogeneous groups results in undesirable trade-offs—between balance, stability, and other metrics—at all equilibria, even though there exists a perfect assessment rule for each group separately.

Decoupling the classifier by group is a natural intervention in this setting. Namely, the institution may choose a group-specific $\theta_a \in \Theta$ to assess individuals from group $a \in \mathcal{A}$, assuming that the group attribute information is available. This corresponds to choosing θ_a that maximizes the utility that the institution derives from each group separately. Thus we now consider the *decoupled dynamics* Φ^{dec} where the institution uses group-specific assessment rules, i.e., for all $a \in \mathcal{A}$

$$\theta_a^{br}(\pi_a) := \underset{\theta_a \in \Theta}{\operatorname{argmax}}\ p_{\text{TP}}\text{TPR}_a(\theta_a)\pi_a - c_{\text{FP}}\text{FPR}_a(\theta_a)(1 - \pi_a).^{6} \quad (5)$$

As in the standard joint setting individuals still acquire qualification according to their group utility as follows

$$\pi_a^{br}(\theta_a) := G(w(\text{TPR}_a(\theta_a) - \text{FPR}_a(\theta_a))).$$

We denote by $\pi^{\text{dec}} \in [0,1]^{|\mathcal{A}|}$ the equilibria of the decoupled dynamics, $\Phi^{\text{dec}} = \left(\pi_a^{br} \circ \theta_a^{br}\right)_{a \in \mathcal{A}}$. It is not hard to see that decoupling is helpful in a group-realizable setting. That is, the qualification rates of the decoupled equilibrium π^{dec} *Pareto-dominates* the qualification rates of all equilibria π under a joint assessment rule, whenever group-realizability holds.

Proposition 4.6 (Decoupling). *Consider a group-realizable setting, that is, for every $a \in \mathcal{A}$, there exists a perfect assessment rule $\theta_a^{\text{opt}} \in \Theta$ such that $\text{TPR}_a(\theta_a^{\text{opt}}) - \text{FPR}_a(\theta_a^{\text{opt}}) = 1$. Then Φ^{dec} has a unique stable equilibrium π^{dec}, where $\pi_a^{\text{dec}} = G(w)$. Moreover, for any equilibrium π of the joint dynamics Φ, $\pi_a^{\text{dec}} \geq \pi_a$ for all $a \in \mathcal{A}$. Furthermore, if there is no perfect assessment rule, i.e.,*

$$\max_{\theta \in \Theta} \sum_{a \in \mathcal{A}} n_a(\text{TPR}_a(\theta) - \text{FPR}_a(\theta)) < 1,$$

then for some $a \in \mathcal{A}$, $\pi_a^{\text{dec}} > \pi_a$.

This proposition directly follows from Proposition 3.1.

Indeed, decoupling always helps in the group-realizable setting—not only does it not decrease any group's equilibrium qualification rate, it also increases the equilibrium qualification rate of at least one group when realizability across all groups does not hold. In Sections 5 and 6 we examine decoupling in the absence of group-realizability and see that those cases are not as clear-cut. When group-realizability does not hold, in some cases decoupling is still helpful while in others it can significantly harm one group.

5 BEYOND GROUP-REALIZABILITY: MULTIPLE EQUILIBRIA WITHIN GROUP

We have so far considered settings where the learning problem is realizable (or almost realizable) within each group. This is a common assumption in various prior works, such as Hu et al. [26]. As we saw in Section 4, there may be multiple undesirable equilibria when a joint assessment rule is used in a group-realizable setting, but these undesirable equilibria disappear in the decoupled dynamics.

In many application domains, realizability does not hold even at a group level. That is to say, no assessment rule in Θ can perfectly separate qualified and unqualified individuals even within one group. This may be due to the fact that mapping individuals to the visible feature space $\mathcal{X}$ involves loss of information or there may be other sources of stochasticity in the domain [12], making it impossible to provide a high accuracy assessment of individuals' qualifications. A key consequence of the lack of realizability is that even for a single group, the optimal classifier now can vary greatly with π_a, the group's qualification rate. As a result, our guarantees

[6] As when we defined the joint dynamics (Section 2), when the argmax is not unique, we assume ties are broken according to a fixed and well-defined order.

about the near-optimality of stable equilibria (Theorem 3.2) no longer hold, and there could exist multiple stable equilibria each corresponding to a different qualification rate within a group. In this section, we investigate the existence of bad equilibria for a single group and its implications on decoupling when the learning problem is not group-realizable. For the rest of this section, we consider a single group, i.e., $|\mathcal{A}| = 1$ and suppress a in the notation.

In the following proposition (proved in Appendix F), we characterize conditions under which multiple equilibria exists for arbitrary feature spaces and assessment rules. This is a generalization of a classical result from Coate and Loury [10] that considers a one-dimensional feature space; we restate and prove the classical result as a consequence of Proposition 5.1 in Appendix F.

Proposition 5.1 (Multiple equilibria in arbitrary feature spaces). *Let Φ be as defined in Section 2. For any qualification rate π, let*

$$\beta(\pi) := \mathrm{TPR}(\theta^{br}(\pi)) - \mathrm{FPR}(\theta^{br}(\pi)),$$

be the difference between true and false positive rates of the institution's utility maximizing assessment rule with respect to π. Assume $\beta(\pi)$ is continuous, the CDF of the cost G is continuous and that there exists $\theta \in \Theta$ such that $\mathbb{P}(\hat{Y}_\theta = 1) = 0$ and $\theta' \in \Theta$ such that $\mathbb{P}(\hat{Y}_{\theta'} = 1) = 1$, i.e., there is a assessment rule that accepts everyone and an assessment rule that rejects everyone. Also suppose the likelihood ratio $\phi(x) := \frac{\mathbb{P}(X=x|Y=0)}{\mathbb{P}(X=x|Y=1)}$ is strictly positive on X.

If $x < G(w\beta(x))$ for some $x \in (0,1)$, then there exists at least two distinct non-zero equilibria where $\pi = \Phi(\pi)$. If in addition β is differentiable, an equilibrium at π is locally stable whenever $G'(w\beta(\pi)) < |\beta'(\pi)|$, where G' and β' denote the derivatives of G and β respectively.

Proposition 5.1 describes conditions under which there exists more than one equilibrium in the dynamics modeled in Section 2. Given a differentiable $\beta(\pi)$, one can always construct a monotonically increasing G, such that the dynamics Φ has any number of locally stable equilibria. The implication of having multiple equilibria is that the dynamics may converge to different equilibrium qualification rates depending on the initial investment, even for a single group. This makes the setting particularly hard to analyze.

Nevertheless, the following result, proved in Appendix F, shows that even in the non-realizable setting, subsidizing the cost of investment by changing the distribution G to a stochastically dominant distribution $\bar{G}$ will create a new equilibrium that has a higher qualification rate. In other words, subsidies in the non-realizable setting also improve the quality of equilibria. However, the new equilibrium is not guaranteed to be locally stable. We see some ramifications of this empirically in the next section.

Proposition 5.2 (Subsidies without realizability). *Suppose $\pi^* > 0$ is an equilibrium for the dynamics Φ_G, where the cost of investment is distributed according to G on $[0,1]$. Let $\bar{G}$ be a CDF that is stochastically dominated by G, that is, $\bar{G}(x) > G(x)$ for all $x \in (0,1)$, and both G and $\bar{G}$ are strictly increasing. Then there exists $\bar{\pi} > \pi^*$ such that $\bar{\pi}$ is an equilibrium for $\Phi_{\bar{G}}$.*

6 SIMULATIONS WITH NON-REALIZABILITY

In this section we present results from numerical experiments examining the effects of decoupling and subsidies under our model of dynamics, in the absence of group-realizability. We consider a stylized semi-synthetic experiment, based on a widely used FICO credit score dataset from a 2007 Federal Reserve report [42]. Importantly, only aggregate statistics were reported and the data we accessed does not contain sensitive or private information. Our modeling assumptions may not be realistic for this dataset (see Section 8) and our simulations should not be interpreted as policy recommendations. Instead, these experiments help us illustrate qualitatively the types of dynamics one may find using real world data.

Stylized Model. We describe how our model can be instantiated to a highly stylized example of credit scoring and lending. Assume a loan applicant either has the means to repay a loan or not. If they have the means to repay, they always repay ($Y = 1$); otherwise they always default ($Y = 0$). In order to have the means to repay, applicants must make an *ex ante* investment at the cost of C, whose distribution is $\mathbb{P}(C < c) = G(c)$. This represents costly actions an individual has to take in order to acquire the financial ability to repay loans, e.g. working at a stable job or taking job preparation classes. Applicants from group a who have the means to repay receive credit scores X drawn from f_1^a and those who don't receive credit scores drawn from f_0^a. The decision of the bank is to approve or reject a loan applicant, given their credit scores.

Dataset. FICO scores are widely used in the United States to predict credit worthiness. The dataset, which contains aggregate statistics, is based on a sample of 301,536 TransUnion TransRisk scores from 2003 [42] and has been preprocessed by Hardt et al. [22]. These scores, corresponding to X in our model, range from 300 to 850. For simplicity, we rescale the scores so that they are between 0 and 1. Individuals were labeled as defaulted if they failed to pay a debt for at least 90 days on at least one account in the ensuing 18-24 month period. The data is also labeled by race, which is the group attribute A that we use. We compute empirical conditional feature distributions $\mathbb{P}(X = x \mid A = a, Y = y)$ from the available data and fit Beta distributions[7] to these to obtain f_0^a, f_1^a.

We treat these distributions *as if* they came from our model as shown in Figure 1, for the sole purpose of illustration. This is not to claim that our modeling assumptions hold on this dataset, as discussed earlier. Given the lending domain is complex, our aim is not to faithfully represent this particular domain with our model, but to simulate feature distributions that exhibit group heterogeneity and non-realizability, hence extending our consideration beyond the idealized settings of Sections 3 and 4.

Figure 3 shows the histograms as well as the fitted Beta distributions for f_0^a, f_1^a, where a is the race attribute. It is clear that group-realizability does not hold even approximately, since there is significant overlap in the distributions of credit scores for people who repaid and for people who did not repay.

6.1 Decoupling and Multiple Stable Equilibria

Although decoupling is guaranteed to improve the qualification rate at equilibrium over using a joint decision rule for every group (Sections 3 and 4), this is not necessarily true in the non-realizable setting. In fact, even when G is the uniform distribution on $[0,1]$ in all groups (i.e. the cost of investment C is uniformly distributed on $[0,1]$, as we considered in Section 4), decoupling did not benefit

[7] We simulate 100,000 samples from the empirical PDF (see Figure 3) and fit a Beta distribution by maximum likelihood estimation.

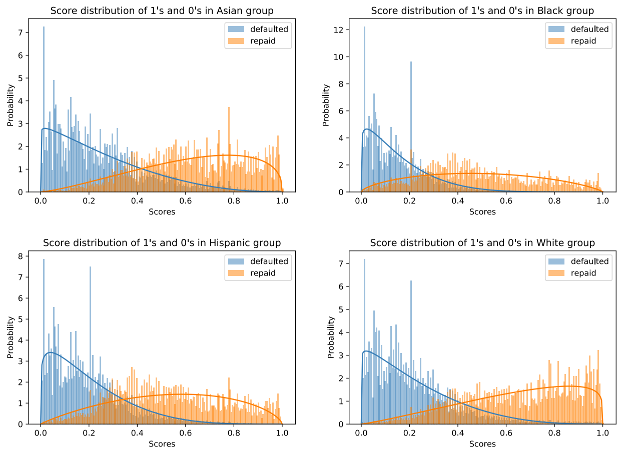

Figure 3: Score distributions conditioning on repayment outcome (Y) for different race groups

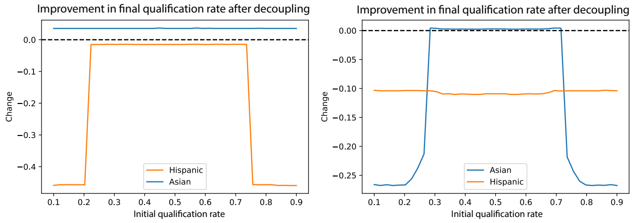

Figure 4: Effects of decoupling in presence of multiple equilibria. We vary the initial qualification rate in the x-axis.

all groups. As can be seen from Figure G.1 in Appendix G, while the White and Asian groups had a higher qualification rate after decoupling, the Black and Hispanic groups saw their equilibrium qualification rate decrease. On the other hand, the effects of decoupling were small in this case (less than 3 percent points difference in the final qualification rate).

We now show that the effect of decoupling can be drastic depending on G. Recall that in Section 5, we showed that multiple equilibria, with possibly vastly different qualification rates, may exist under the non-realizable setting even when there is a only single group. In general the existence of multiple equilibria depends on properties of G, that is, how the cost of investment is distributed in a group. In Figure 4, we show the change in equilibrium investment level after decoupling for an experiment with two groups, Asian and Hispanic. The two plots each correspond to a different bimodal Gaussian distribution for G, truncated to $[0, 1]$, that have been chosen such that the decoupled dynamics have multiple stable equilibria for the Hispanic (right) and the Asian (left) respectively[8].

In both plots, we can see that the effects of decoupling depend on the initial qualification rate. If the initial qualification rate was too low, or too high, the decoupled dynamics converge to an equilibrium where one of the groups invest in qualifications at a much lower level than they would under the joint dynamics.[9]

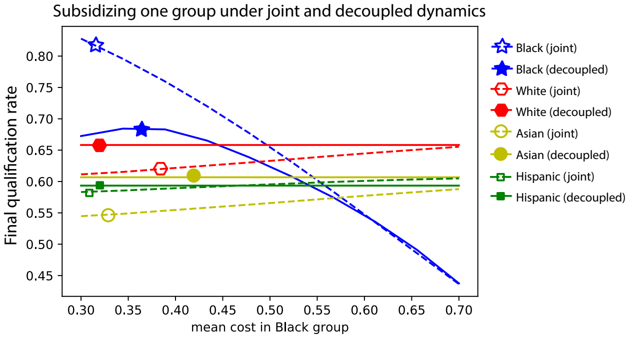

Figure 5: Effects of raising the average cost of investment, by varying the mean of G on the x-axis.

6.2 Subsidizing the Cost of Investment

In this experiment, we consider if subsidizing the cost of investment of one group by changing G improves their new equilibrium qualification rate, under both decoupled and joint dynamics. Specifically, we vary the cost of investment in the Black group.

We use a truncated normal distribution for G and vary its mean (on the x-axis) for a single group, while keeping the other groups' G unchanged (mean of 0.6).

Figure 5 shows that subsidizing the cost of investment is effective in raising the equilibrium investment level of a group, both in the joint learning and decoupled learning case. Interestingly, large amounts of subsidy for a single group reduced the equilibrium investment levels of other groups. As also suggested by theoretical results in section 4.3, subsidizing the qualification rate of one group does sometimes entail a tradeoff in the qualification rates of other possibly more advantaged groups.

Interestingly, lowering the mean cost of investment in the Black group below 0.35 caused the final qualification rate to decrease. This is not a contradiction of Proposition 5.2, which argues that equilibria improve under subsidies but does not guarantee that the dynamics will converge to the improved equilibrium. In this case, the decoupled dynamics for the Black group (where the mean cost of investment is 0.30) actually converged to a limit cycle and the final qualification rate in the plot is an average of the points in the limit cycle. Limit cycles are a challenging object to study in dynamical systems and game theory. While we have commented on their existence in a simple model in group-realizable setting of Section 4.2, we leave their implications in the general non-realizable setting to future work.

7 RELATED WORK

Our work follows a growing line of work on how machine learning algorithms interact with human actors in a dynamic setting, with the goal of understanding and mitigating disparate impact.

Recent work examine the long-term impact of group fairness criteria [see e.g., 4, Chapter 2] on automated decision making systems: Liu et al. [33] show that static fairness criteria fail to account for the delayed impact that decisions have on the welfare of disadvantaged groups. In the context of hiring, however, Hu and Chen [25] find that applying the demographic parity constraint in a temporary labor market achieves an equitable long-term equilibrium in the permanent labor market by raising worker reputations.

[8] The right (resp. left) plot is generated using a bimodal normal distribution for G with modes at 0.57 and 0.74 (resp. at 0.57 and 0.63).

[9] See Figure G.2 in Appendix G for the converged qualification rates of both groups.

Prior work on the fairness of machine learning has examined tradeoffs between fairness criteria [9, 32], as well as the incompatibility between risk minimization and fairness criteria [34], assuming that the qualification rates differ across groups. These results concern the static setting, whereas we highlight the fact that qualification rates tend to change in response to the decision rules.

Another line of work [23, 45] analyzes a dynamic model where users respond to errors made by an institution by leaving the user base uniformly at random, and demonstrate how the risk-minimizing approach to machine learning can amplify representation disparity over time. This is complementary to our work which models individuals as rational decision makers who may or may not have the incentive to acquire the positive label. In particular, Hashimoto et al. [23] show that equilibria with equal user representation from all groups can be unstable, and that robust learning can stabilize such equilibria. Unlike in our model, the user representation model does not distinguish between positive and negative labels, and thus do not distinguish between false negative and false positive errors. This is a crucial distinction in high-stakes decision making as different error types present asymmetric incentives for individuals, as explained in Section 2; for example, a high false positive rate in hiring would encourage under-qualified job applicants.

Hu et al. [26] and Milli et al. [36] study the disparate impact of being robust towards strategic manipulation [see e.g., 21], where individuals respond to machine learning systems by manipulating their features to get a better classification. In contrast to our model (Figure 5), their setting models the individual as intervening directly on their features, X, and this is assumed to have no effect on their qualification Y. This assumption applies to features that are easy to 'game' (e.g. scores on standardized tests can be improved by test preparation classes) but is less applicable to features that more directly correspond to *investment* in one's qualifications (e.g. taking AP courses in high school). Hu et al. [26] also show that subsidizing the costs of the disadvantaged group to strategically manipulate their features can sometimes lead to harmful effects. Kleinberg and Raghavan [31] and Khajehnejad et al. [29] study decision policies that incentivize individuals to directly manipulate their features X to optimize particular notions of utility.

Our work is also related to the topic of statistical discrimination in economics [2, 3, 40], which studies how disparate market outcomes at equilibrium can arise from imperfect information. This line of work often involves wage discrimination, whereas we assume the wage is fixed and standard for all groups. Coate and Loury [10] proposed a model of rational individual investment in the labor market under a fixed wage and showed that affirmative action may lead to an undesirable equilibrium where one group still invests sub-optimally. The model in our work is most closely related to their model, with two key distinctions: 1) We allow features X to be multi-dimensional, whereas Coate and Loury [10] assumes that X is a one-dimensional 'noisy signal'. 2) We consider the case where the conditional feature distributions, $\mathbb{P}(X = x \mid Y = y, A = a)$, differ by groups whereas Coate and Loury [10] assumes that the groups are identically distributed. Under our models, if the conditional feature distributions were shared across groups, then *any* hiring policy will result in fully balanced equilibria where all groups have the same qualification rate and are hired at the same rate. This does not corroborate with reality, where conditional feature distributions do in fact differ across groups and we routinely observe institutions applying the same model to all individuals only to see obviously discriminatory outcomes [14]. By modeling feature heterogeneity across groups, we find it necessarily leads to disparate equilibria.

Recently, Mouzannar et al. [37] studied the equilibria of qualification rates under a generic class of dynamics, focusing on contractive maps and the effects of affirmative action. In contrast, our work motivates a model of dynamics based on rational investment, and this typically leads to non-contractive dynamics. We are both interested in balanced equilibria, which they termed 'social equality'.

Finally, our work studies two interventions for finding more desirable equilibria: decoupling the classifier and subsidizing the cost of investment. Several works, including Dwork et al. [17] and Ustun et al. [43], have studied decoupled classifiers in the static classification setting. Our work sheds light on when such interventions are useful in the dynamic decision making setting.

8 DISCUSSION AND FUTURE WORK

In this work, we have made the following contributions:

(1) We proposed a dynamic model of decision making where individuals invest rationally based on the current assessment rule. We studied the properties of equilibria under these dynamics.
(2) We showed that common properties of real data, namely heterogeneity across groups and the lack of realizability, lead to undesirable tradeoffs at equilibria, resulting in long term outcomes that disadvantage one or more groups.
(3) We considered two interventions—decoupling and subsidizing the cost of investment—and showed that they have a significant impact on the nature of equilibria both in theory and in numerical experiments.

We now discuss the limitations of the current work and avenues for future research. Questions related to sampling and its ramifications for the nature of equilibria are challenging and warrant further study. This work assumed that the institution can estimate the true and false positive rates over the entire population, even though it really can only observe the qualification of candidates *after* hiring them. This is known as the selective labeling problem, which could introduce bias. In theory, unbiased estimates can be achieved by a small degree of random sampling and appropriate reweighting [see e.g., 30], but this is still a large problem in practice that requires domain-specific knowledge and solutions [15, 28].

Our model assumed that individuals make a rational decision to invest and can affect their qualification Y directly. This assumption could be reasonable in settings like hiring, for example, where investing to acquire skills usually leads to increased competence. In some settings, however, individuals may be unable to effectively intervene on Y. For example, a business loan applicant who is a good business operator could still default on their loan due to external economic shocks or other forms of disadvantage that have not been taken into account. In this case, the current model does not fully capture the complex societal processes that lead to a positive outcome. Our work nonetheless shows that even in an idealized setting where individuals can effectively and rationally intervene on their outcome labels Y, underlying factors such as heterogeneity across groups and non-realizability already lead to undesirable tradeoffs at equilibrium. We leave the extensions of the current model beyond rational individual investment to future work.

REFERENCES

[1] Abhay P. Aneja and Carlos F. Avenancio-Leon. 2019. No Credit For Time Served? Incarceration and Credit-Driven Crime Cycles.

[2] Kenneth J. Arrow. 1973. The Theory of Discrimination. In *Discrimination in Labor Markets*. Princeton University Press, 3–33.

[3] Kenneth J. Arrow. 1998. What Has Economics to Say about Racial Discrimination? *Journal of Economic Perspectives* 12, 2 (Spring 1998), 91–100.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[5] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 671 (2016).

[6] David Card and Jesse Rothstein. 2007. Racial segregation and the black-ĂŞwhite test score gap. *Journal of Public Economics* 91, 11 (2007), 2158 – 2184.

[7] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM, New York, NY, USA, 224–232. https://doi.org/10.1145/3240323.3240370

[8] Raj Chetty, Nathaniel Hendren, and Lawrence F. Katz. 2016. The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. *American Economic Review* 106, 4 (2016), 855–902.

[9] A. Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5 (2017), 153–163. Issue 2.

[10] Stephen Coate and Glenn C. Loury. 1993. Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review* 83, 5 (1993), 1220–1240.

[11] Michael Conover, Jacob Ratkiewicz, Matthew R Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. *ICWSM* 133 (2011), 89Ð96.

[12] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *CoRR* abs/1808.00023 (2018).

[13] Kate Crawford. 2017. The Trouble with Bias. NeurIPS Keynote.

[14] Jeffrey Dastin. 2019. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (10 2019).

[15] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. 2018. Learning under selective labels in the presence of expert consistency. *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)* (2018).

[16] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 120–128. https://doi.org/10.1145/3287560.3287572

[17] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research)*, Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 119–133.

[18] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*. 160–171.

[19] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* 55, 1 (1997), 119–139.

[20] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2017. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *Technical report, CEPR Discussion Papers* (2017).

[21] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science (ITCS '16)*. ACM, New York, NY, USA, 111–122.

[22] M. Hardt, E. Price, and N. Srebo. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*. 3315–3323.

[23] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholm, Sweden, 1929–1938.

[24] Caroline Hoxby and Christopher Avery. 2013. The Missing "One-Off": The Hidden Supply of High-Achieving, Low-Income Students. *Brookings Papers on Economic Activity* 1 (2013), 1–65.

[25] Lily Hu and Yiling Chen. 2018. A Short-term Intervention for Long-term Fairness in the Labor Market. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1389–1398.

[26] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 259–268. https://doi.org/10.1145/3287560.3287597

[27] Ehud Kalai. 2004. Large Robust Games. *Econometrica* 72, 6 (2004), 1631–1665.

[28] Nathan Kallus and Angela Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 2439–2448.

[29] Moein Khajehnejad, Behzad Tabibian, Bernhard Schölkopf, Adish Singla, and Manuel Gomez-Rodriguez. 2019. Optimal Decision Making Under Strategic Behavior. *CoRR* abs/1905.09239 (2019). arXiv:1905.09239

[30] Niki Kilbertus, Manuel Gomez-Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2019. Improving Consequential Decision Making under Imperfect Predictions. *CoRR* abs/1902.02979 (2019). arXiv:1902.02979

[31] Jon Kleinberg and Manish Raghavan. 2019. How Do Classifiers Induce Agents to Invest Effort Strategically?. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC '19)*. ACM, New York, NY, USA, 825–844.

[32] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. [n. d.]. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* ([n. d.]).

[33] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholm, Sweden, 3150–3158.

[34] Lydia T. Liu, Max Simchowitz, and Moritz Hardt. 2019. The Implicit Fairness Criterion of Unconstrained Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Vol. 97. PMLR, Long Beach, California, USA, 4051–4060.

[35] Stella Lowry and Gordon Macpherson. 1988. A blot on the profession. *British Medical Journal* 296, 6623 (1988), 657–658.

[36] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. The Social Cost of Strategic Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 230–239.

[37] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. 2019. From Fair Decision Making To Social Equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 359–368.

[38] Eli Pariser. 2011. *The Filter bubble: What the Internet is hiding from you.* Penguin, UK.

[39] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York, NY, USA.

[40] Edmund Phelps. 1972. The Statistical Theory of Racism and Sexism. *American Economic Review* 62 (02 1972), 659–61.

[41] David Schmeidler. 1973. Equilibrium points of nonatomic games. *Journal of Statistical Physics* 7, 4 (01 Apr 1973), 295–300.

[42] US Federal Reserve. 2007. Report to the congress on credit scoring and its effects on the availability and affordability of credit.

[43] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without Harm: Decoupled Classifiers with Preference Guarantees. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 6373–6382.

[44] Meredith Whittaker, Kate Crawford, Genevieve Fried Roel Dobbe, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. AI Now Report 2018.

[45] Xueru Zhang, Mohammad Mahdi Khalili, Cem Tekin, and Mingyan Liu. 2019. Long term impact of fair machine learning in sequential decision making: representation disparity and group retention. *CoRR* abs/1905.00569 (2019). arXiv:1905.00569