Fairness Interventions as (Dis)Incentives for Strategic Manipulation

Xueru Zhang[1], **Mohammad Mahdi Khalili**[2], **Kun Jin**[3], **Parinaz Naghizadeh**[4], **Mingyan Liu**[3]

[1]Computer Science and Engineering, The Ohio State University
[2]Computer and Information Sciences, University of Delaware
[3] Electrical and Computer Engineering, University of Michigan
[4]Integrated Systems Engineering & Electrical and Computer Engineering, The Ohio State University
zhang.12807@osu.edu, khalili@udel.edu, kunj@umich.edu, naghizadeh.1@osu.edu, mingyan@umich.edu

Abstract

Although machine learning (ML) algorithms are widely used to make decisions about individuals in various domains, concerns have arisen that (1) these algorithms are vulnerable to strategic manipulation and "gaming the algorithm"; and (2) ML decisions may exhibit bias against certain social groups. Existing works have largely examined these as two separate issues, e.g., by focusing on building ML algorithms robust to strategic manipulation, or on training a fair ML algorithm. In this study, we set out to understand the impact they each have on the other, and examine how to design fair algorithms in the presence of strategic behavior. The strategic interaction between a decision maker and individuals (as decision takers) is modeled as a two-stage (Stackelberg) game; when designing an algorithm, the former anticipates the latter may manipulate their features in order to receive more favorable decisions. We analytically characterize the equilibrium strategies of both, and examine how the algorithms and their resulting fairness properties are affected when the decision maker is strategic (anticipates manipulation), as well as the impact of fairness interventions on equilibrium strategies. In particular, we identify conditions under which anticipation of strategic behavior may mitigate/exacerbate unfairness, and conditions under which fairness interventions can serve as incentives/disincentives for strategic manipulation.

1 Introduction

As machine learning (ML) algorithms are increasingly being used to make high-stake decisions in domains such as hiring, lending, criminal justice, and college admissions, the need for transparency increases in terms of how decisions are reached given input. However, given (partial) information about an algorithm, individuals subject to its decisions can and will adapt their behavior by strategically manipulating their data in order to obtain favorable decisions. This strategic behavior in turn hurts the performance of ML models and diminishes their utility. Such a phenomenon has been widely observed in real-world applications, and is known as *Goodhart's law*, which states "once a measure becomes a target, it ceases to be a good measure" (Strathern 1997). For instance, a hiring or admissions practice that heavily depends on GPA might motivate students to cheat on exams; not accounting for such manipulation may result in disproportionate hiring of under-qualified individuals. A strategic decision maker is one who anticipates such behavior and thus aims to make its ML models robust to such strategic manipulation.

A second challenge facing ML algorithms is the growing concern over bias in their decisions, and various notions of fairness (e.g., demographic parity (Barocas, Hardt, and Narayanan 2019), equal opportunity (Hardt, Price, and Srebro 2016)) have been proposed to measure and remedy biases. These measures typically impose an (approximate) equality constraint over certain statistical measures (e.g., positive classification rate, true positive rate, etc.) across different groups when building ML algorithms.

In this paper, we study the design of (fair) machine learning algorithms in the presence of strategic manipulation. Specifically, we consider a decision maker whose goal is to select individuals that are *qualified* for certain tasks based on a given set of features. Given knowledge of the selection policy, individuals can tailor their behavior and manipulate their features in order to receive favorable decisions. We shall assume that this feature manipulation does not affect an individual's true qualification state. We say the decision maker (and its policy) is *strategic* if it anticipates such manipulation; it is *non-strategic* if it does not take into account individuals' manipulation in its policies.

We adopt a typical two-stage (Stackelberg) game setting where the decision maker commits to its policies, following which individuals best-respond. A crucial difference between this study and existing models of strategic interaction is that existing models typically assume features and their manipulation are deterministic so that the manipulation cost can be modeled as a function of the change in features (Hardt et al. 2016; Dong et al. 2018; Milli et al. 2019; Hu, Immorlica, and Vaughan 2019; Braverman and Garg 2020; Brückner and Scheffer 2011; Haghtalab et al. 2020; Kleinberg and Raghavan 2019; Chen, Wang, and Liu 2020; Miller, Milli, and Hardt 2020); by contrast, in our setting features are random variables whose realizations are *unknown* prior to an individual's manipulation decision. In fact, this is the case in many important applications, a motivating example is presented in Sec. 2.

Moreover, among these existing works, only (Milli et al. 2019; Hu, Immorlica, and Vaughan 2019; Braverman and Garg 2020) studied the disparate impact of ML decisions on different social groups, where the disparity stems from different manipulation costs and different feature distributions. No

fairness intervention was considered in these works. In contrast, we study the impact of fairness intervention on different groups in the presence of strategic manipulative behavior, and explore the role of fairness intervention in (dis)incentivizing such manipulation. We aim to answer the following questions: how does the anticipation of individuals' strategic behavior impact a decision maker's utility, and the resulting policies' fairness properties? How is the Stackelberg equilibrium affected when fairness constraints are imposed? Can fairness intervention serve as (dis)incentives for individuals' manipulation? More related work is discussed in Appendix A.

Our main contributions and findings are as follows.

1. We formulate a Stackelberg game to model the interaction between a decision maker and strategic individuals (Sec. 2). We characterize both strategic (fair) and non-strategic (fair) optimal policies of the decision maker, and individuals' best response (Sec. 3, Lemmas 1-4).
2. We study the impact of the decision maker's anticipation of individuals' strategic manipulation by comparing non-strategic with strategic policies (Sec. 4):
   - We show that compared to non-strategic policy, strategic policy always disincentivizes manipulative behavior; it over (resp. under) selects when a population is majority-qualified (resp. majority-unqualified)[1](Thm. 1).
   - We show that the anticipation of manipulation can *worsen* the fairness of a strategic policy: when one group is majority-qualified while the other is majority-unqualified (Thm. 2); on the other hand, when both groups are majority-unqualified, we show the possibility of using strategic policy to *mitigate unfairness* and even flip the disadvantaged group (Thm. 3).
3. We study the impact of fairness interventions on policies and individuals' manipulation (Sec. 5).
   - If a decision maker lacks information or awareness to anticipate manipulative behavior (but which in fact exists), we identify conditions under which such non-strategic decision maker benefits from using fairness constrained policies rather than unconstrained policies (Thm. 4).
   - By comparing individuals' responses to a strategic policy with and without fairness intervention, we identify scenarios under which a strategic fair policy can (dis)incentivize manipulation compared to an unconstrained strategic policy (Thm. 5 and Thm. 6).
4. We examine our theoretical findings using both synthetic and real-world datasets (Sec. 7).

2 Problem Formulation

Consider two demographic groups $\mathcal{G}_a$, $\mathcal{G}_b$ distinguished by a sensitive attribute $S \in \{a, b\}$ (e.g., gender), with fractions $n_s = \Pr(S = s)$ of the population. An individual from either group has observable features $X \in \mathbb{R}^d$ and a hidden qualification state $Y \in \{0, 1\}$. Let $\alpha_s = P_{Y|S}(1|s)$ be the qualification rate of $\mathcal{G}_s$. A decision maker makes a decision $D \in \{0, 1\}$ ("0" being negative/reject and "1" positive/accept) for an individual using a group-dependent policy $\pi_s(x) = P_{D|XS}(1|x, s)$. An individual's action is denoted by $M \in \{0, 1\}$, with $M = 1$ indicating manipulation and $M = 0$ otherwise. Note that in our context manipulation does not change the true qualification state Y. It is the qualification state Y, sensitive attribute S, and manipulation action M together that drive the realizations of features X.

Best response. An individual in $\mathcal{G}_s$ incurs a random cost $C_s \geq 0$ when manipulating its features, with probability density function (PDF) $f_s(c)$ and cumulative density function (CDF) $\mathbb{F}_{C_s}(c) = \int_0^c f_s(z)dz$. The realization of this random cost is known to an individual when determining its action M; the decision maker on the other hand only knows the overall cost distribution of each group. Thus the response that the decision maker anticipates (from the group as a whole or from a randomly selected individual) is expressed as follows, whereby given policy π_s, an individual in $\mathcal{G}_s$ will manipulate its features if doing so increases its utility:

$$wP_{D|YMS}(1|y, 1, s) - C_s \geq wP_{D|YMS}(1|y, 0, s).$$

Here $w > 0$ is a fixed benefit to the individual associated with a positive decision $D = 1$ (the benefit is 0 otherwise); without loss of generality we will let $w = 1$. In other words, the best response the decision maker expects from the individuals of $\mathcal{G}_s$ with qualification y is their probability of manipulation, denoted by p_s^y and written as:

$$p_s^y(\pi_s) = \Pr\big(C_s \leq P_{D|YMS}(1|y, 1, s) - P_{D|YMS}(1|y, 0, s)\big).$$

We assume that individuals manipulate by imitating the features of those qualified, e.g., students cheat on exams by hiring a qualified person to take exams (or copying answers of those qualified), job applicants manipulate resumes by mimicking those of the skilled employees, loan applicants fool the lender by using/stealing identities of qualified people, etc. This is inspired by the *imitative learning* behavior observed in social learning, whereby new behaviors are acquired by copying social models' actions (Ganos et al. 2012; Gergely and Csibra 2006). Under this assumption, the qualified individuals do not have incentives to manipulate (as manipulation doesn't bring additional benefit but cost) and only those unqualified may choose to manipulate, i.e., $P_{M|YS}(1|1, s) = 0$. To simplify the notations, we will use $P_{X|YS}(x|y, s)$ to denote the distributions *before* manipulation. The feature distribution of those unqualified after manipulation becomes $(1 - p_s^0(\pi_s))P_{X|YS}(x|0, s) + p_s^0(\pi_s)P_{X|YS}(x|1, s)$.

Motivating Example: The above formulation is fundamentally different from existing literature: 1) we consider uncertain manipulated outcomes where individuals only have probabilistic knowledge of how features may change upon manipulation; 2) the realization of manipulation cost is fixed and known to each individual, as opposed to being a function of features before and after manipulation. A prime example is students cheating on an exam by paying for someone else to take it, where the exam score is treated as feature (in making admissions or employment decisions): (i) here individual's own score and the manipulated feature outcome (actual score received upon hiring an imposter) are random, but individuals have a good idea from past experience what those score distributions would be like; (ii) the cost of hiring someone is

[1] A group is majority-(un)qualified if the majority of that population is (un)qualified.

more or less fixed, and it is determined by the outcome (the fake score) rather than the difference in score improvement. As the real test score was never realized (students who hire someone actually never take the exam themselves), there is really no way to compute precisely how much the feature has improved and put a price on it even after the fact. The existing model does not fit such applications.

Optimal (fair) policy. The decision maker receives a true-positive (resp. false-positive) benefit (resp. penalty) u_+ (resp. u_-) when accepting a qualified (resp. unqualified) individual. Its utility, denoted by $R(D, Y)$, is $R(1, 1) = u_+$, $R(1, 0) = u_-$, $R(0, 0) = R(0, 1) = 0$. The decision maker aims to find optimal policies for the two groups such that its expected total utility $\mathbb{E}[R(D, Y)]$ is maximized.

As mentioned earlier, there are two types of decision makers, strategic and non-strategic: A *strategic decision maker* anticipates strategic manipulation, has perfect information on the manipulation cost distribution, and accounts for this in determining policies, while a *non-strategic decision maker* ignores manipulative behavior in determining its policies. Either type may further impose a fairness constraint $\mathcal{C}$, to ensure that π_a and π_b satisfy the following:

$$\mathbb{E}_{X\sim\mathcal{P}_a^{\mathcal{C}}}[\pi_a(X)] = \mathbb{E}_{X\sim\mathcal{P}_b^{\mathcal{C}}}[\pi_b(X)]\,, \tag{1}$$

where $\mathcal{P}_s^{\mathcal{C}}$ is some probability distribution over X associated with fairness constraint $\mathcal{C}$. Many fairness notions can be written in this form, e.g., equal opportunity (`EqOpt`) (Hardt, Price, and Srebro 2016) where $\mathcal{P}_s^{\texttt{EqOpt}}(x) = P_{X|YS}(x|1, s)$, or demographic parity (`DP`) (Barocas, Hardt, and Narayanan 2019) where $\mathcal{P}_s^{\texttt{DP}}(x) = P_{X|S}(x|s)$.

The above leads to four types of optimal policies a decision maker can use, which we consider in this paper: 1) a non-strategic policy; 2) a non-strategic fair policy; 3) a strategic policy; 4) a strategic fair policy. These are detailed in Sec. 3.

The Stackelberg game. The interaction between the decision maker and individuals consists of the following two stages in sequence: (i) The former publishes its policies (π_a, π_b), which may be strategic or non-strategic, and may or may not satisfy a fairness constraint, and (ii) the latter, while observing the published policies and their realized costs, decide whether to manipulate their features.

3 Four types of (non-)strategic (fair) policies

Non-strategic policy. A decision maker who does not account for individuals' strategic manipulation maximizes the expected utility $\widehat{U}_s(\pi_s)$ over $\mathcal{G}_s$ defined as follows:

$$\int_X \Big[u_+\alpha_s P_{X|YS}(x|1, s) - u_-(1-\alpha_s)P_{X|YS}(x|0, s)\Big]\pi_s(x)dx.$$

Define $\mathcal{G}_s$'s *qualification profile* as $\gamma_s(x) = P_{Y|XS}(1|x, s)$. Then, we can show that the non-strategic policy $\widehat{\pi}_s^{\texttt{UN}} = \text{argmax}_{\pi_s}\widehat{U}_s(\pi_s)$ is in the form of a threshold policy, i.e., $\widehat{\pi}_s^{\texttt{UN}}(x) = \mathbf{1}\big(\gamma_s(x) \geq \frac{u_-}{u_+ + u_-}\big)$ (Appendix G). Throughout the paper, we will present results in the one dimensional feature space. Generalization to high dimensional spaces is discussed in Appendix B.

Assumption 1. *$P_{X|YS}(x|1, s)$, $P_{X|YS}(x|0, s)$ are continuous and satisfy the strict monotone likelihood ratio property, i.e., $\frac{P_{X|YS}(x|1,s)}{P_{X|YS}(x|0,s)}$ is increasing in $x \in \mathbb{R}$. Let unique x_s^* be such that $P_{X|YS}(x_s^*|1, s) = P_{X|YS}(x_s^*|0, s)$.*

Assumption 1 is relatively mild and can be satisfied by distributions such as exponential and Gaussian, and has been widely used (Zhang et al. 2020; Jung et al. 2020; Barman and Rathi 2020; Khalili et al. 2021; Coate and Loury 1993). It implies that an individual is more likely to be qualified as their feature value increases. Under Assumption 1, the threshold policy can be written as $\pi_s(x) = \mathbf{1}(x \geq \theta_s)$ for some $\theta_s \in \mathbb{R}$. Throughout the paper, we assume Assumption 1 holds and focus on threshold policies. We will frequently use θ_s to denote policy π_s. Under Assumption 1, the thresholds for non-strategic policies are characterized as follows.

Lemma 1. *Let $(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})$ be non-strategic optimal thresholds. Then $\frac{P_{X|YS}(\widehat{\theta}_s^{UN}|1,s)}{P_{X|YS}(\widehat{\theta}_s^{UN}|0,s)} = \frac{u_-(1-\alpha_s)}{u_+\alpha_s}$.*

Non-strategic fair policy. Denoted as $(\widehat{\pi}_a^{\mathcal{C}}, \widehat{\pi}_b^{\mathcal{C}})$, this is found by maximizing the total utility subject to fairness constraint $\mathcal{C}$, i.e., $(\widehat{\pi}_a^{\mathcal{C}}, \widehat{\pi}_b^{\mathcal{C}}) = \text{argmax}_{(\pi_a,\pi_b)} n_a\widehat{U}_a(\pi_a) + n_b\widehat{U}_b(\pi_b)$ such that Eqn (1) holds. It can be shown that for `EqOpt` and `DP` fairness, the optimal fair policies are also threshold policies and can be characterized by the following (Zhang et al. 2020).

Lemma 2 ((Zhang et al. 2020))**.** *Let $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ be thresholds in non-strategic optimal fair policies. These satisfy*

$$\sum_{s=a,b} n_s \left(\frac{u_+\alpha_s P_{X|YS}(\widehat{\theta}_s^{\mathcal{C}}|1, s) - u_-(1-\alpha_s)P_{X|YS}(\widehat{\theta}_s^{\mathcal{C}}|0, s)}{\mathcal{P}_s^{\mathcal{C}}(\widehat{\theta}_s^{\mathcal{C}})}\right) = 0.$$

Strategic policy. Let $p_s^0 := P_{M|YS}(1|0, s)$, the probability that unqualified individuals in $\mathcal{G}_s$ manipulate. Under policy $\pi_s(x) = \mathbf{1}(x \geq \theta)$, the decision maker's expected utility $U_s(\theta)$ over $\mathcal{G}_s$ is as follows:

$$\widehat{U}_s(\theta) - u_-(1-\alpha_s)\Big(\mathbb{F}_{X|YS}(\theta|0, s) - \mathbb{F}_{X|YS}(\theta|1, s)\Big)p_s^0$$

where $\widehat{U}_s(\theta)$ is the expected utility under non-strategic policy, $\mathbb{F}_{X|YS}(x|y, s) = \int_{-\infty}^{x} P_{X|YS}(z|y, s)dz$ denotes the CDF.

Define *manipulation benefit* as

$$\Delta_s(\theta) := \mathbb{F}_{X|YS}(\theta|0, s) - \mathbb{F}_{X|YS}(\theta|1, s),$$

representing the additional benefit an individual gains from manipulation. Then, the unqualified individuals' best-response (i.e., manipulation probability introduced in Sec. 2) to policy $\pi_s(x) = \mathbf{1}(x \geq \theta)$ can be equivalently written as

$$p_s^0(\theta) := p_s^0(\pi_s) = \mathbb{F}_{C_s}(\Delta_s(\theta)).$$

The detailed derivation is in Appendix G. This manipulation probability $p_s^0(\theta)$ is single-peaked with maximum occurring at x_s^*, and $\lim_{\theta\to-\infty} p_s^0(\theta) = \lim_{\theta\to+\infty} p_s^0(\theta) = 0$, meaning that when the threshold is sufficiently low or high, unqualified individuals are less likely to manipulate their features. Plugging this in the decision maker's utility, we have

$$U_s(\theta) = \widehat{U}_s(\theta) - \underbrace{u_-(1-\alpha_s)\Delta_s(\theta)\mathbb{F}_{C_s}(\Delta_s(\theta))}_{\textbf{term 2}:=\Psi_s(\Delta_s(\theta))}. \tag{2}$$

Define a function $\Psi_s(z) := u_-(1-\alpha_s)\mathbb{F}_{C_s}(z)z$, then **term 2** in Eqn. (2) can be written as $\Psi_s(\Delta_s(\theta))$, and can be interpreted as the additional loss incurred by the decision maker due to manipulation (equivalently, the average manipulation gain by group $\mathcal{G}_s$). Further, let $\Psi'_s(z)$ be denoted as the first order derivative of $\Psi_s(z)$, then $\Psi'_s(\Delta_s(\theta))$ indicates the decision maker's *marginal loss* caused by strategic manipulation (equivalently, the *marginal* manipulation gain of $\mathcal{G}_s$). The thresholds for strategic policies are characterized as follows.

Lemma 3. *For $(\theta_a^{UN}, \theta_b^{UN})$, the strategic optimal thresholds, $\frac{P_{X|YS}(\theta_s^{UN}|1,s)}{P_{X|YS}(\theta_s^{UN}|0,s)} = \frac{u_-(1-\alpha_s)-\Psi'_s(\Delta_s(\theta_s^{UN}))}{u_+\alpha_s-\Psi'_s(\Delta_s(\theta_s^{UN}))}$.*

Strategic fair policy. Strategic fair thresholds $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ are found by maximizing the total expected utility subject to fairness constraint $\mathcal{C}$, i.e., $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}}) = \text{argmax}_{(\theta_a,\theta_b)} n_a U_a(\theta_a) + n_b U_b(\theta_b)$ such that Eqn. (1) holds. They can be characterized by the following.

Lemma 4. *Let $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$ be thresholds in strategic optimal fair policies. These satisfy*

$$\sum_{s=a,b} n_s \Big(\frac{P_{X|YS}(\theta_s^{\mathcal{C}}|0,s) - P_{X|YS}(\theta_s^{\mathcal{C}}|1,s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s^{\mathcal{C}})} \Psi'_s(\Delta_s(\theta_s^{\mathcal{C}})) + \frac{u_+\alpha_s P_{X|YS}(\theta_s^{\mathcal{C}}|1,s) - u_-(1-\alpha_s)P_{X|YS}(\theta_s^{\mathcal{C}}|0,s)}{\mathcal{P}_s^{\mathcal{C}}(\theta_s^{\mathcal{C}})} \Big) = 0.$$

Note that besides $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$, the equations in Lemmas 3 and 4 may be satisfied by other threshold pairs that are not optimal. We discuss this further in the next section.

4 Impact of anticipating manipulations

Impact on the optimal policy & utility function. We first compare strategic policy θ_s^{UN} with non-strategic policy $\widehat{\theta}_s^{\text{UN}}$, and examine how the policy and the decision maker's expected utility differ. Let $\overline{\Delta}_s := \max_\theta \Delta_s(\theta)$.

Assumption 2. *$\Psi'_s(z) < \infty$ is non-decreasing over $[0, \overline{\Delta}_s]$.*

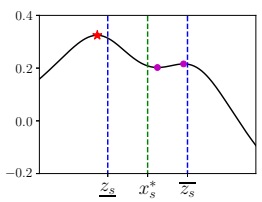
For any threshold θ, $\Delta_s(\theta)$ represents the manipulation benefit of $\mathcal{G}_s$; those in $\mathcal{G}_s$ choose to manipulate if $C_s \le \Delta_s(\theta)$. Therefore, $\overline{\Delta}_s$ indicates the maximum additional benefit an individual in $\mathcal{G}_s$ may gain from manipulation. As $\Psi'_s(\Delta_s(\theta))$ represents the marginal manipulation gain of $\mathcal{G}_s$ on average, Assumption 2 means that a group's *marginal* manipulation gain does not decrease as manipulation benefit increases. Examples (e.g., beta/uniformly distributed cost) satisfying this assumption can be found in Appendix C. Note that under Assumption 2, $\Psi'_s(0) = 0$ and $\Psi'_s(\Delta_s(\theta))$ is single-peaked with maximum occurring at x_s^*. We assume it holds in Sections 4 and 5. Define $\nu_s = \max\{u_+\alpha_s, u_-(1-\alpha_s)\}$.

Theorem 1. *Let $\overline{\Psi'_s} = \Psi'_s(\overline{\Delta}_s)$, $\delta_u = \frac{u_-}{u_-+u_+}$, and $\underline{z_s} < \overline{z_s}$ be defined such that $\Psi'_s(\Delta_s(\underline{z_s})) = \Psi'_s(\Delta_s(\overline{z_s})) = \nu_s$.*

1. *If $\alpha_s = \delta_u$, then $\theta_s^{UN} = \widehat{\theta}_s^{UN} = x_s^*$ when $\overline{\Psi'_s} \le \nu_s$, and $\theta_s^{UN} \in \{\underline{z_s}, \overline{z_s}\}$ otherwise.*
2. *If $\alpha_s < \delta_u$ (resp. $\alpha_s > \delta_u$), then $\theta_s^{UN} > \widehat{\theta}_s^{UN} > x_s^*$ (resp. $\theta_s^{UN} < \widehat{\theta}_s^{UN} < x_s^*$). Moreover, if $\overline{\Psi'_s} > \nu_s$, then $\widehat{\theta}_s^{UN} > \overline{z_s}$ (resp. $\widehat{\theta}_s^{UN} < \underline{z_s}$) and $U_s(\theta)$ may have additional extreme points in $(\underline{z_s}, x_s^*)$ (resp. $(x_s^*, \overline{z_s})$); otherwise $\widehat{\theta}_s^{UN}$ is the unique extreme point of $U_s(\theta)$.*

Note that although $\widehat{U}_s(\theta)$ (non-strategic utility) and $\Psi_s(\Delta_s(\theta))$ are single-peaked with unique extreme points, their difference $U_s(\theta)$ (Eqn.(2)) may have multiple extreme points. As we will see later, this results in strategic and non-strategic policies having different properties in many aspects.

An example of $U_s(\theta)$ is shown to the right: $X|Y = y, S = s \sim \mathcal{N}(\mu^y, 4.7^2)$, $[\mu^0, \mu^1] = [-5, 5]$, $C_s \sim \text{Beta}(10, 4)$, $\alpha_s = 0.6$ and $u_- = u_+$. The red star is the optimal threshold $\theta_s^{\text{UN}} < \underline{z_s}$; two magenta dots are other extreme points of $U_s(\theta)$, which are in $(x_s^*, \overline{z_s})$. Theorem 1 states that $U_s(\theta)$ has multiple extreme points if $\overline{\Psi'_s}$ is sufficiently large, and it also specifies the range of those extreme points.



Note that the maximum marginal manipulation gain $\overline{\Psi'_s}$ depends on $P_{X|YS}(x|y,s)$, α_s, and C_s. Given fixed cost C_s, $\overline{\Psi'_s}$ increases as the maximum manipulation benefit $\overline{\Delta}_s$ increases and/or α_s decreases (i.e., when there are more unqualified individuals who can manipulate). Given fixed $\overline{\Delta}_s$ and α_s, $\overline{\Psi'_s}$ increases as cost decreases (i.e., $f_s(c)$ is shifted/skewed toward the direction of lower cost). Theorem 1 shows that as compared to non-strategic policy $\widehat{\theta}_s^{\text{UN}}$, strategic policy θ_s^{UN} over(under) selects when a group is majority-(un)qualified.[2] In either case, as shown by Theorem 1, this means $\widehat{\theta}_s^{\text{UN}}$ is always closer to x_s^* (the single peak of $p_s^0(\theta)$) compared to θ_s^{UN}. Therefore, the strategic policy always disincentivizes manipulative behavior, i.e., manipulation probability $p_s^0(\theta_s^{\text{UN}}) < p_s^0(\widehat{\theta}_s^{\text{UN}})$.

Impact on fairness. The characterization of strategic policy $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and non-strategic policy $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ allows us to further compare them against a given fairness criterion $\mathcal{C}$. Suppose we define the *unfairness* of threshold policy (θ_a, θ_b) as $\mathcal{E}^{\mathcal{C}}(\theta_a, \theta_b) = \mathbb{E}_{X\sim\mathcal{P}_a^{\mathcal{C}}}[\mathbf{1}(x \ge \theta_a)] - \mathbb{E}_{X\sim\mathcal{P}_b^{\mathcal{C}}}[\mathbf{1}(x \ge \theta_b)] = \mathbb{F}_b^{\mathcal{C}}(\theta_b) - \mathbb{F}_a^{\mathcal{C}}(\theta_a)$, where the CDF $\mathbb{F}_s^{\mathcal{C}}(\theta) = \int_{-\infty}^{\theta} \mathcal{P}_s^{\mathcal{C}}(x)dx$. Define the *disadvantaged group* under policy (θ_a, θ_b) as the group with the larger $\mathbb{F}_s^{\mathcal{C}}(\theta_s)$, i.e., the group with the smaller selection rate (DP) or the smaller true positive rate (EqOpt). Define group index $-s := \{a, b\} \setminus s$. Note that we measure unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a, \theta_b)$ over the original feature distributions $P_{X|YS}(x|y,s)$ before manipulation.

We first identify distributional conditions under which the strategic optimal policy worsens unfairness.

Theorem 2. *If $\alpha_s > \delta_u > \alpha_{-s}$ and $\mathbb{F}_s^{\mathcal{C}}(x_s^*) \le \mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^*)$, then strategic policy $(\theta_a^{UN}, \theta_b^{UN})$ has worse unfairness compared to non-strategic $(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})$, i.e., $\left|\mathcal{E}^{\mathcal{C}}(\theta_a^{UN}, \theta_b^{UN})\right| > \left|\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})\right|$, $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$. Moreover, the disadvantaged group under $(\theta_a^{UN}, \theta_b^{UN})$ and $(\widehat{\theta}_a^{UN}, \widehat{\theta}_b^{UN})$ is the same.*

Given the conditions in Thm. 2, $\mathcal{G}_{-s}$ is disadvantaged under non-strategic policy. Because the majority-(un)qualified

[2] We say $\mathcal{G}_s$ is majority-unqualified (resp. majority-qualified) if $\alpha_s < \delta_u$ (resp. $\alpha_s > \delta_u$). When $u_- = u_+$, a group is majority-(un)qualified if more than a half of its members are (un)qualified.

group $\mathcal{G}_s(\mathcal{G}_{-s})$ is over(under) selected under strategic policy (Theorem 1), $\mathcal{G}_{-s}$ becomes more disadvantaged while $\mathcal{G}_s$ becomes more advantaged, i.e., the unfairness gap is wider under strategic policy. Note that condition $\mathbb{F}^{\mathcal{C}}_s(x^*_s) \leq \mathbb{F}^{\mathcal{C}}_{-s}(x^*_{-s})$ holds if $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$. For the DP fairness measure, it holds for any distribution when α_s is sufficiently large or α_{-s} sufficiently small. As shown in Sec. 7, it is also seen in the real world (e.g., FICO data).

We next identify conditions on the manipulation cost, under which strategic policy $(\theta^{\text{UN}}_a, \theta^{\text{UN}}_b)$ can lead to a more equitable outcome or flip the (dis)advantaged group compared to non-strategic $(\widehat{\theta}^{\text{UN}}_a, \widehat{\theta}^{\text{UN}}_b)$.

Theorem 3. *If $\alpha_a, \alpha_b < \delta_u$ and $\mathbb{F}^{\mathcal{C}}_{-s}(\widehat{\theta}^{UN}_{-s}) > \mathbb{F}^{\mathcal{C}}_s(\widehat{\theta}^{UN}_s)$, i.e., $\mathcal{G}_{-s}$ is disadvantaged under non-strategic policy, then given any $\mathcal{G}_{-s}$, there always exists cost C_s for $\mathcal{G}_s$ such that $\overline{\Psi'_s}$ is sufficiently large and*

1. *$(\theta^{UN}_a, \theta^{UN}_b)$ mitigates the unfairness; or*
2. *$(\theta^{UN}_a, \theta^{UN}_b)$ flips the disadvantaged group from $\mathcal{G}_{-s}$ to $\mathcal{G}_s$.*

Because $\alpha_s < \delta_u$, we have $\theta^{\text{UN}}_s > \widehat{\theta}^{\text{UN}}_s > x^*_s$ (by Thm. 1). Moreover, θ^{UN}_s increases as $\Psi'_s(\Delta_s(\theta))$ increases ($f_s(c)$ is skewed toward the direction of lower cost). Intuitively, as $\mathcal{G}_s$'s manipulation cost decreases, more individuals can afford manipulation; thus a strategic decision maker disincentivizes manipulation by increasing the threshold θ^{UN}_s. For any $\mathcal{G}_{-s}$, as $\mathbb{F}^{\mathcal{C}}_s(\theta^{\text{UN}}_s)$ increases, either the unfairness gets mitigated or $\mathbb{F}^{\mathcal{C}}_s(\theta^{\text{UN}}_s)$ becomes larger than $\mathbb{F}^{\mathcal{C}}_{-s}(\theta^{\text{UN}}_{-s})$. Proposition 1 in Appendix E considers a special case when $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$, and gives conditions on $\Psi'_s(\cdot)$ under which $(\theta^{\text{UN}}_a, \theta^{\text{UN}}_b)$ mitigates the unfairness or flips the disadvantaged group when $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$.

5 Impact of fairness interventions

In this section, we study how non-strategic and strategic policies are affected by fairness interventions $\mathcal{C} \in \{\texttt{DP}, \texttt{EqOpt}\}$.

Impact of fairness intervention on non-strategic policy. First, we consider a non-strategic decision maker and compare $(\widehat{\theta}^{\text{UN}}_a, \widehat{\theta}^{\text{UN}}_b)$ with $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$, both ignoring strategic manipulation but the latter imposing a fairness criterion. Theorem 4 identifies conditions under which a fairness constrained $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$ yields *higher* utility from both groups compared to unconstrained $(\widehat{\theta}^{\text{UN}}_a, \widehat{\theta}^{\text{UN}}_b)$. It is worth noting because had strategic manipulation been absent, $(\widehat{\theta}^{\text{UN}}_a, \widehat{\theta}^{\text{UN}}_b)$ by definition would attain the optimal/highest utility for decision maker.

Theorem 4. *Let $\nu_s = \max\{u_+\alpha_s, u_-(1-\alpha_s)\}$, suppose $\Psi'_b(\Delta_b(\widehat{\theta}^{\mathcal{C}}_b)) > \nu_b$ and $\Psi'_a(\Delta_a(\widehat{\theta}^{\mathcal{C}}_a)) > \nu_a$. When $\mathbb{F}^{\mathcal{C}}_s(\widehat{\theta}^{UN}_s) < \mathbb{F}^{\mathcal{C}}_{-s}(\widehat{\theta}^{UN}_{-s})$ (i.e., $\mathcal{G}_{-s}$ is disadvantaged under non-strategic optimal policy), $U_a(\widehat{\theta}^{\mathcal{C}}_a) > U_a(\widehat{\theta}^{UN}_a)$ and $U_b(\widehat{\theta}^{\mathcal{C}}_b) > U_b(\widehat{\theta}^{UN}_b)$ hold under any of the following cases: 1) $\alpha_s < \delta_u < \alpha_{-s}$; 2) $\alpha_a, \alpha_b > \delta_u$ and $\alpha_s \to \delta_u$; 3) $\alpha_a, \alpha_b < \delta_u$ and $\alpha_{-s} \to \delta_u$.*

Condition $\alpha_s, \alpha_{-s} \to \delta_u$ means that the qualification rates α_s, α_{-s} are sufficiently close to δ_u. Thm. 4 says that when the marginal manipulation gains of the groups under non-strategic fair policy $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$ are sufficiently large, $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$ may outperform $(\widehat{\theta}^{\text{UN}}_a, \widehat{\theta}^{\text{UN}}_b)$ in terms of both fairness and utility due to the misalignment of $U_s(\theta)$ and $\widehat{U}_s(\theta)$ caused by manipulation. This means that if the decision maker lacks information or awareness to anticipate manipulative behavior (but which in fact exists), then it would benefit from using a fairness constrained policy $(\widehat{\theta}^{\mathcal{C}}_a, \widehat{\theta}^{\mathcal{C}}_b)$ rather than $(\widehat{\theta}^{\text{UN}}_a, \widehat{\theta}^{\text{UN}}_b)$.

Impact of fairness intervention on the strategic policy. We now compare $(\theta^{\text{UN}}_a, \theta^{\text{UN}}_b)$ with $(\theta^{\mathcal{C}}_a, \theta^{\mathcal{C}}_b)$. We also explore their respective subsequent impact on individuals' manipulative behavior by comparing manipulation probabilities $\big(p^0_a(\theta^{\text{UN}}_a), p^0_b(\theta^{\text{UN}}_b)\big)$ and $\big(p^0_a(\theta^{\mathcal{C}}_a), p^0_b(\theta^{\mathcal{C}}_b)\big)$. The goal here is to understand whether fairness intervention can serve as incentives or disincentives for strategic manipulation. According to Thm. 1, $U_s(\theta)$ may have multiple extreme points under strategic manipulation if the group's marginal manipulation gain is sufficiently large. Depending on whether $U_s(\theta)$ has multiple extreme points, different conclusions result as outlined in Thm. 5 below, which identifies conditions under which fairness intervention may increase the manipulation incentive for one group while disincentivizing the other, or it may serve as incentives for both groups.

Theorem 5 (Fairness as (dis)incentives)**.** *Denote $p^{\mathcal{C}}_s := p^0_s(\theta^{\mathcal{C}}_s)$ and $p^{UN}_s := p^0_s(\theta^{UN}_s)$, we have:*

1. *When both $U_a(\theta)$ and $U_b(\theta)$ have unique extreme points, then $\theta^{UN}_s > \theta^{\mathcal{C}}_s$ and $\theta^{UN}_{-s} < \theta^{\mathcal{C}}_{-s}$ must hold. Moreover,*
   - *i) If $\alpha_s > \delta_u > \alpha_{-s}$, then $\forall \alpha_{-s}$, $\exists \kappa, \tau \in (0,1)$ such that $\forall \alpha_s > \kappa$ and $\forall n_s > \tau$, we have $p^{UN}_s < p^{\mathcal{C}}_s$, $p^{UN}_{-s} > p^{\mathcal{C}}_{-s}$.*
   - *ii) If $\alpha_a, \alpha_b > \delta_u$ (resp. $\alpha_a, \alpha_b < \delta_u$), then $\forall \alpha_{-s}$, there exists $\kappa \in (\delta_u, 1)$ (resp. $\kappa \in (0, \delta_u)$) such that $\forall\ \alpha_s > \kappa$ (resp. $\alpha_s < \kappa$), we have $(p^{UN}_a - p^{\mathcal{C}}_a)(p^{UN}_b - p^{\mathcal{C}}_b) < 0$.*
2. *When at least one of $U_a(\theta)$, $U_b(\theta)$ has multiple extreme points, then it is possible that $\forall s \in \{a, b\}$, $\theta^{UN}_s > \theta^{\mathcal{C}}_s$ or $\theta^{UN}_s < \theta^{\mathcal{C}}_s$, i.e., both groups are over/under selected under fair policies. In this case,*
   - *i) If $\alpha_s > \delta_u > \alpha_{-s}$, then $(p^{UN}_s - p^{\mathcal{C}}_s)(p^{UN}_{-s} - p^{\mathcal{C}}_{-s}) < 0$.*
   - *ii) If $\alpha_a, \alpha_b > \delta_u$ (or $\alpha_a, \alpha_b < \delta_u$), then either $p^{UN}_a < p^{\mathcal{C}}_a$, $p^{UN}_b < p^{\mathcal{C}}_b$ or $(p^{UN}_a - p^{\mathcal{C}}_a)(p^{UN}_b - p^{\mathcal{C}}_b) < 0$.*

When not accounting for strategic manipulation, $\widehat{U}_s(\theta)$ has a unique extreme point, and imposing a fairness constraint results in one group getting under-selected and the other over-selected. In contrast, when the decision maker anticipates strategic manipulation, $U_s(\theta)$ may have multiple extreme points. One consequence of this difference is that both $\mathcal{G}_a$ and $\mathcal{G}_b$ may be over- or under-selected when fairness is imposed, resulting in more complex incentive relationships. Specifically, if one group is majority-qualified while the other is majority-unqualified, then under-selecting (resp. over-selecting) both groups under fair policies will increase (resp. decrease) the incentives of the former to manipulate, while disincentivizing (resp. incentivizing) the latter (by *2.(i)*); if both groups are majority-(un)qualified, then the fair policy may incentivize both to manipulate (by *2.(ii)*).

If the marginal manipulation gain of both groups are not sufficiently large, i.e., $U_s(\theta)$ has a unique extreme point, then fairness intervention always results in one group getting over-selected and the other under-selected. However,

its subsequent impact on incentives may vary depending on $P_{X|YS}(x|y,s)$, n_s. Thm. 5 identifies two scenarios under which fair policies incentivize one group (say $\mathcal{G}_s$) while disincentivizing the other ($\mathcal{G}_{-s}$): when $\mathcal{G}_s$ is majority-qualified, $\mathcal{G}_{-s}$ majority-unqualified, and $\mathcal{G}_s$ sufficiently qualified ($\alpha_s > \kappa$) and represented in the entire population ($n_s > \tau$) (by *1.(i)*); or, when both are majority-(un)qualified and $\mathcal{G}_s$ sufficiently (un)qualified (by *1.(ii)*).

Next, we identify conditions under which fairness intervention can *disincentivize* both groups. Let x_s^{UN} be defined s.t. $\Delta_s(x_s^{\text{UN}}) = \Delta_s(\theta_s^{\text{UN}})$ and $x_s^{\text{UN}} \neq \theta_s^{\text{UN}}$ when $\theta_s^{\text{UN}} \neq x_s^*$. Note that x_s^{UN} is the point at which $p_s^0(x_s^{\text{UN}}) = p_s^0(\theta_s^{\text{UN}})$. Because manipulation probability is single-peaked, fairness intervention incentivizes manipulative behavior of $\mathcal{G}_s$ if $\theta_s^{\mathcal{C}}$ falls between x_s^{UN} and θ_s^{UN}.

Theorem 6 (Disincentives for both groups)**.** *Suppose $U_a(\theta)$ and $U_b(\theta)$ have unique extreme points. If $\alpha_a, \alpha_b > \delta_u$ (resp. $\alpha_a, \alpha_b < \delta_u$) and $\mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{UN}) < \mathbb{F}_s^{\mathcal{C}}(x_s^*)$ (resp. $\mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{UN}) > \mathbb{F}_s^{\mathcal{C}}(x_s^*)$), then $\exists \kappa, \tau \in (0,1)$ s.t. $\forall \alpha_s \in (\delta_u, \kappa)$ (resp. $\alpha_s \in (\kappa, \delta_u)$) and $\forall n_s > \tau$, we have $p_a^{UN} > p_a^{\mathcal{C}}$ and $p_b^{UN} > p_b^{\mathcal{C}}$.*

Note that x_s^* depends on $P_{X|YS}(x|y,s)$ and x_{-s}^{UN} is determined by u_-, u_+, $P_{X|YS}(x|y,-s)$ and α_{-s}. Thm. 6 says that when both groups are majority-(un)qualified, for certain population distributions and $\mathcal{G}_{-s}$, fair policies disincentivize both groups if $\mathcal{G}_s$ is sufficiently unqualified(qualified) and sufficiently represented in the population. For a special Gaussian case, conditions for satisfying $\mathbb{F}_{-s}^{\mathcal{C}}(x_{-s}^{\text{UN}}) \lesseqgtr \mathbb{F}_s^{\mathcal{C}}(x_s^*)$ in Thm. 6 are given in Proposition 2 in Appendix E.

Theorems 5 and 6 suggest that the impact of fairness intervention on the individuals' manipulative behavior highly depends on manipulation costs, feature distributions, group qualification and representation. This complexity stems from the misalignment in manipulation probability $p_s^0(\theta)$, utility $U_s(\theta)$, and fairness $\mathcal{C}$. In particular, the manipulation probability of $\mathcal{G}_s$ is single-peaked with maximum at x_s^*, which does not depend on group qualification and representation, but on which the decision maker's total utility depends, as varying α_s and n_s will affect the policies. As a result, depending on which region θ_s^{UN} falls into, i.e., smaller or larger than x_s^*, and how it may change under constraint $\mathcal{C}$, fairness intervention will have different impacts on incentives.

Although Theorems 5 and 6 hold for both `EqOpt` and `DP` fairness, there are scenarios under which they have different impact on incentives. Proposition 3 in Appendix E considers a special case when $P_{X|YS}(x|y,a) = P_{X|YS}(x|y,b)$ and one group is majority-qualified while the other majority-unqualified, in which `EqOpt` never disincentivize both groups while `DP` can disincentivize both.

6 Discussion

In practice, individual strategic behavior can be much more complicated than modeled here: those considered qualified may also have an incentive to manipulate, and manipulation may only lead to partial improvement in features. The latter can be modeled by introducing a sequence of progressively "better" distributions (each with a different manipulation cost), and the goal of manipulation is to imitate/acquire a distribution better than one's own. The model studied in this paper is essentially the two-distribution (one for the unqualified, one for the qualified) version of this more general model. Even in this simplified model, there exists a complex relationship between fairness intervention and incentives for strategic manipulation as we have shown. Our results provide insights and build a foundation for analyzing more complicated models in future work.

Our present model is limited to scenarios where individual qualification states and manipulation actions are binary. In reality, qualification states can be on a continuous spectrum, and individuals may face more complex manipulation decisions such as what features to manipulate, what types of actions to take, etc., than a binary decision of whether to manipulate or not. Going beyond the binary settings is also a direction of future research.

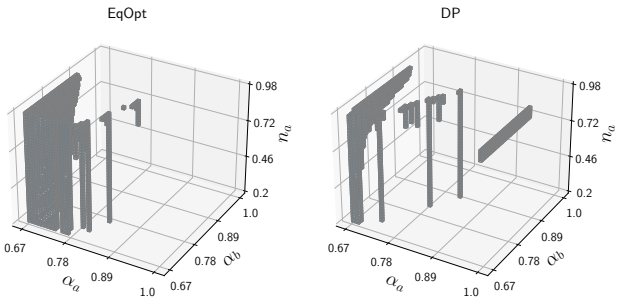


Fig. 1: $\alpha_s, \alpha_b > \delta_u$, $C_a = C_b \sim \text{Beta}(10,1)$, $\frac{u_+}{u_-} = \frac{1}{2}$. Grey region is $(\alpha_a, \alpha_b, n_a)$ satisfying $\mathbb{F}_b^{\mathcal{C}}(x_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(x_a^*)$ in Thm. 6; meanwhile both groups are disincentivized under $(\theta_a^{\mathcal{C}}, \theta_b^{\mathcal{C}})$.

7 Experiments

We conduct experiments on both a Gaussian synthetic dataset, and the FICO scores dataset (Reserve 2007). Due to the lack of real-world data on manipulation cost, we consider manipulation costs following either uniform ($C_s \sim U[0, \bar{c}]$) or beta distributions ($C_s \sim \text{Beta}(a,b)$, smaller b and larger a lead to larger manipulation costs, see Fig. 7 in Appendix F).[3] Note that these are examples for illustration, our results do not rely on these choices.

Gaussian data. Suppose $X|Y=y, S=s$ is Gaussian distributed. Fig. 1 shows an example where fairness intervention can serve as disincentive for manipulation for both groups. It shows $\forall \alpha_b > \delta_u$ satisfying condition $\mathbb{F}_b^{\mathcal{C}}(x_b^{\text{UN}}) < \mathbb{F}_a^{\mathcal{C}}(x_a^*)$, there exist sufficiently small α_a and sufficiently large n_a under which $p_a^0(\theta_a^{\text{UN}}) > p_a^0(\theta_a^{\mathcal{C}})$ and $p_b^0(\theta_b^{\text{UN}}) > p_b^0(\theta_b^{\mathcal{C}})$, i.e., both groups are disincentivized under strategic fair policy. This verifies Thm. 6. Detailed parameters and more experiments (e.g., verification of Theorems 2, 3, and 5) on Gaussian data can be found in Appendix F.

FICO scores (Reserve 2007). FICO scores are widely used in the US to assess an individual's creditworthiness. The is a dataset pre-processed by (Hardt, Price, and Srebro 2016)

[3] Uniformly distributed C_s has been adopted in (Liu et al. 2020). In economics, a choice of *generalized beta distribution* is common to model costs (e.g., healthcare costs (Jones, Lomas, and Rice 2014)).

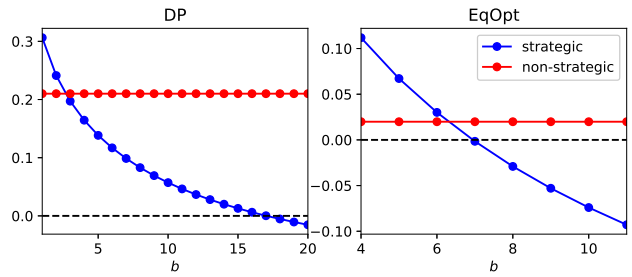


Fig. 2: Unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$, $\frac{u_+}{u_-} = \frac{1}{2}$, $\alpha_a, \alpha_b < \delta_u$. Perfect equity is indicated by the black dashed line. $C_b \sim \text{Beta}(10, 5)$, $C_a \sim \text{Beta}(10, b)$, where larger b indicates smaller costs.

Table 1: Unfairness $\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and $\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ for $\mathcal{C} \in \{\texttt{EqOpt}, \texttt{DP}\}$: $\mathcal{G}_b$ = African-American, $u_+ = u_-$, $C_a \sim \text{Beta}(10, 2)$. When cost $C_a \neq C_b$, $C_b \sim \text{Beta}(10, 6)$.

	$\mathcal{G}_a$	strategic $C_a = C_b$	strategic $C_a \neq C_b$	non-strategic
EqOpt	Caucasian	0.355	0.556	0.136
	Hispanic	0.292	0.493	0.034
	Asian	0.333	0.533	0.123
DP	Caucasian	0.611	0.680	0.449
	Hispanic	0.421	0.490	0.242
	Asian	0.634	0.703	0.522

to generate CDF of scores $\mathbb{F}_{X|S}(x|s)$ and qualification profile $P_{Y|XS}(1|x, s)$ for different social groups (Caucasian, African-American, Hispanic, Asian). We use these to estimate the conditional feature distribution $P_{X|YS}(x|y, s)$ by fitting the simulated data to a Beta distribution. This allows us to derive the various equilibrium strategies studied in this paper. We further calculate repayment rates α_s and proportions n_s. These are summarized in Figs. 14 & 15 and Table 3 in Appendix F. Here we focus on beta distributed costs, results for the uniformly distributed C_a, C_b are in Appendix F.

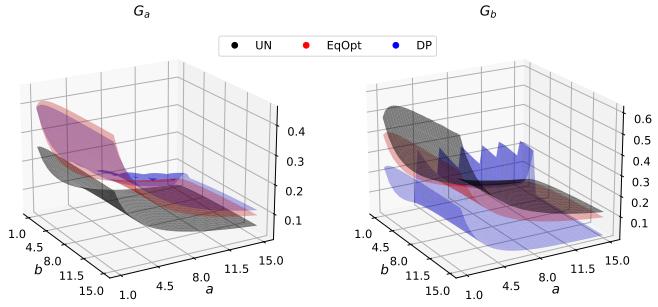
We first compare strategic $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$ and non-strategic policy $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$ in terms of their fairness. Let $\mathcal{G}_a$ denote Caucasian, Hispanic or Asian, and $\mathcal{G}_b$ denote African-American. As shown in Table 1, $\mathcal{G}_b$ is always disadvantaged compared to other groups, and strategic policy worsens unfairness. When $C_a \neq C_b$, the manipulation cost of $\mathcal{G}_b$ is shifted lower. It further shows that this gets worse when it is less costly for those in $\mathcal{G}_b$ to manipulate their features. Since $\alpha_a > \delta_u > \alpha_b$, this is consistent with Thm. 2.

Fig. 2 illustrates how unfairness can be mitigated and how the disadvantaged group can gain advantage under strategic policy. Specifically, let $\mathcal{G}_a, \mathcal{G}_b$ be Hispanic and African-American respectively. We fix $\mathcal{G}_b$ and vary $\mathcal{G}_a$'s manipulation cost. It shows while $\mathcal{G}_b$ is disadvantaged under non-strategic policy ($\mathcal{E}^{\mathcal{C}}(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}}) > 0$), unfairness can be mitigated under strategic policy as $\mathcal{G}_a$'s manipulation cost decreases, and the disadvantaged group may gain an advantage in the process ($\mathcal{E}^{\mathcal{C}}(\theta_a^{\text{UN}}, \theta_b^{\text{UN}}) < 0$). This is an example of Thm. 3.

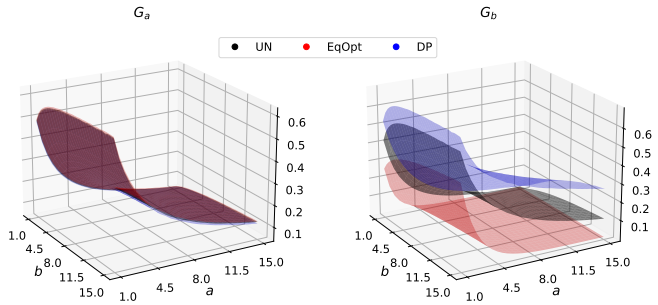
According to Thm. 4, under strategic manipulation, non-strategic fair policy $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ may yield higher utilities from both groups compared to $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$. We verify this in Table 2, in which $\mathcal{G}_a$, $\mathcal{G}_b$ denote Caucasian and Asian respectively, with EqOpt as the fairness constraint. It illustrates two cases corresponding to cases 1 and 2 in Thm. 4, and $U_a(\widehat{\theta}_a^{\mathcal{C}}) > U_a(\widehat{\theta}_a^{\text{UN}})$, $U_b(\widehat{\theta}_b^{\mathcal{C}}) > U_b(\widehat{\theta}_b^{\text{UN}})$ hold in both cases, i.e., $(\widehat{\theta}_a^{\mathcal{C}}, \widehat{\theta}_b^{\mathcal{C}})$ satisfies fairness and attains higher utility than $(\widehat{\theta}_a^{\text{UN}}, \widehat{\theta}_b^{\text{UN}})$.

Table 2: $\mathcal{G}_a$ = Caucasian($\alpha_a = 0.758$), $\mathcal{G}_b$ = Asian($\alpha_b = 0.804$), $\mathcal{C}$ = EqOpt. $C_b \sim \text{Beta}(10, 10)$. The first (resp. second) row corresponds to case 1 (resp. case 2) in Thm. 4.

δ_u	C_a	$U_a(\widehat{\theta}_a^{\text{UN}})$	$U_a(\widehat{\theta}_a^{\mathcal{C}})$	$U_b(\widehat{\theta}_b^{\text{UN}})$	$U_b(\widehat{\theta}_b^{\mathcal{C}})$
0.8	Beta(10, 10)	-0.190	-0.189	0.024	0.034
0.756	Beta(10, 1)	0.396	0.397	0.181	0.201



(a) $\mathcal{G}_a$ = African-American; $\mathcal{G}_b$ =Hispanic



(b) $\mathcal{G}_a$ = Caucasian; $\mathcal{G}_b$ =Asian

Fig. 3: Manipulation probabilities under strategic (fair) policy: $C_a = C_b \sim \text{Beta}(a, b)$, $a, b \in [1, 15]$.

Lastly, we examine how fairness intervention acts as incentives for manipulation. Manipulation probabilities $p_s^0(\theta_s^{\text{UN}})$, $p_s^0(\theta_s^{\texttt{EqOpt}})$, and $p_s^0(\theta_s^{\texttt{DP}})$ are compared under different manipulation costs in Fig. 3. Here groups have the same manipulation costs $C_a = C_b \sim \text{Beta}(a, b)$ and $u_- = u_+$. Experiments on different manipulation costs ($C_a \sim U[0, \bar{c}_a]$, $C_b \sim U[0, \bar{c}_b]$) are shown in Appendix F. Black, red and blue surfaces indicate the manipulation probabilities $p_s^0(\theta_s)$ under $(\theta_a^{\text{UN}}, \theta_b^{\text{UN}})$, $(\theta_a^{\texttt{EqOpt}}, \theta_b^{\texttt{EqOpt}})$ and $(\theta_a^{\texttt{DP}}, \theta_b^{\texttt{DP}})$ policies as manipulation costs change. It shows that fairness intervention can incentivize both groups to manipulate (Fig. 3a), and that EqOpt and DP may have contrarian impact (Fig. 3b). More experiments on other group pairs are in Appendix F.

References

Alon, T.; Dobson, M.; Procaccia, A.; Talgam-Cohen, I.; and Tucker-Foltz, J. 2020. Multiagent evaluation mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1774–1781.

Barman, S.; and Rathi, N. 2020. Fair Cake Division Under Monotone Likelihood Ratios. In *Proceedings of the 21st ACM Conference on Economics and Computation*, 401–437.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bechavod, Y.; Ligett, K.; Wu, S.; and Ziani, J. 2021. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, 1234–1242. PMLR.

Braverman, M.; and Garg, S. 2020. The Role of Randomness and Noise in Strategic Classification. In *1st Symposium on Foundations of Responsible Computing*.

Brückner, M.; Kanzow, C.; and Scheffer, T. 2012. Static prediction games for adversarial learning problems. *The Journal of Machine Learning Research*, 13(1): 2617–2654.

Brückner, M.; and Scheffer, T. 2011. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 547–555.

Chen, Y.; Wang, J.; and Liu, Y. 2020. Strategic Recourse in Linear Classification. *arXiv preprint arXiv:2011.00355*.

Coate, S.; and Loury, G. C. 1993. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 1220–1240.

Dong, J.; Roth, A.; Schutzman, Z.; Waggoner, B.; and Wu, Z. S. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 55–70.

Ganos, C.; Ogrzal, T.; Schnitzler, A.; and Münchau, A. 2012. The pathophysiology of echopraxia/echolalia: relevance to Gilles de la Tourette syndrome. *Movement Disorders*, 27(10): 1222–1229.

Gergely, G.; and Csibra, G. 2006. Sylvia's recipe: The role of imitation and pedagogy in the transmission of cultural knowledge. *Roots of human sociality: Culture, cognition, and human interaction*, 229–255.

Haghtalab, N.; Immorlica, N.; Lucier, B.; and Wang, J. Z. 2020. Maximizing Welfare with Incentive-Aware Evaluation Mechanisms. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 160–166.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

Hu, L.; Immorlica, N.; and Vaughan, J. W. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268.

Jones, A. M.; Lomas, J.; and Rice, N. 2014. Applying beta-type size distributions to healthcare cost regressions. *Journal of Applied Econometrics*, 29(4): 649–670.

Jung, C.; Kannan, S.; Lee, C.; Pai, M.; Roth, A.; and Vohra, R. 2020. Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, 677–678.

Khalili, M. M.; Zhang, X.; Abroshan, M.; and Sojoudi, S. 2021. Improving Fairness and Privacy in Selection Problems. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kleinberg, J.; and Raghavan, M. 2019. How Do Classifiers Induce Agents to Invest Effort Strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 825–844.

Liu, L. T.; Wilson, A.; Haghtalab, N.; Kalai, A. T.; Borgs, C.; and Chayes, J. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 381–391.

Miller, J.; Milli, S.; and Hardt, M. 2020. Strategic Classification is Causal Modeling in Disguise. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6917–6926. PMLR.

Milli, S.; Miller, J.; Dragan, A. D.; and Hardt, M. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 230–239.

Reserve, U. F. 2007. 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*.

Shavit, Y.; Edelman, B.; and Axelrod, B. 2020. Causal strategic linear regression. In *International Conference on Machine Learning*, 8676–8686. PMLR.

Strathern, M. 1997. 'Improving ratings': audit in the British University system. *European review*, 5(3): 305–321.

Sundaram, R.; Vullikanti, A.; Xu, H.; and Yao, F. 2021. PAC-Learning for Strategic Classification. In *International Conference on Machine Learning*, 9978–9988. PMLR.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 18457–18469.