Could Anticipating Gaming Incentivize Improvement in (Fair) Strategic Classification?

Sura Alhanouti and Parinaz Naghizadeh

***Abstract*— As machine learning algorithms increasingly influence crucial decisions in areas like loan approvals and hiring, understanding human strategic behavior in response to these systems becomes vital. We explore strategic manipulation and improvement actions by individuals facing algorithmic decisions, the algorithm designer's role in shaping these strategic responses, and the fairness implications. We formulate these interactions as a Stackelberg game, where a firm deploys a (fair) classifier, and individuals strategically respond. Unlike previous research, our model incorporates both different costs and stochastic efficacy for manipulation and improvement. The analysis reveals different potential classes of agent responses, and characterizes optimal classifiers. Based on these, we highlight the impact of the firm's anticipation of strategic behavior, identifying cases when a (fair) strategic policy can motivate improvement while reducing manipulation.**

I. Introduction

Machine learning (ML) algorithms have come to play a pivotal role in guiding decision making in many application areas, including banking, hiring, social media, and resource allocation. While the use of ML-driven systems can enhance efficiency, it can also drive the humans who are subject to algorithmic decisions to adjust their behavior accordingly. Examples include Uber drivers coordinating their behavior in response to its surge pricing algorithm [1], applicants selecting keywords and formatting to pass automated resume screening [2], and Facebook [3] users adjusting their posting and content interaction choices in response to the platforms' curation algorithms. These can be viewed as *strategic* responses by rational human subjects in these systems, motivating a game-theoretical analysis of learning algorithms with human in the loop.

Recent research has considered settings where agents strategically manipulate observable data (features) to secure favorable outcomes, engaging in what is known as *strategic manipulation* [4]–[7]. However, the alternative option of *strategic improvement*, wherein agents invest genuine effort to modify their true qualification states and attain favorable results, remains less studied [8]–[12]; further, the fairness implications of the availability of both types of strategic behavior remains unexplored.

This paper addresses this gap by analyzing a binary classification problem where agents face a choice between manipulation (solely changing their features) and improvement (investing effort which leads to changes in both their features and true qualification states). Our work differs from prior literature in that our model incorporates not only distinct costs for these strategic actions, but also different (stochastic) efficacy for them, and further has both costs and efficacy varying across demographic groups. This allows us to unravel the effects of these differences on agents' decisions, providing insights that firms can leverage to adapt their algorithms in a way that encourages improvements while discouraging manipulations.

Formulating the problem as a Stackelberg game, the firm first deploys a (fair) classifier, and then agents strategically respond to it. We characterize the Stackelberg game equilibrium under the assumption that improvement is more costly than manipulation, but that it is more effective at advancing agents' chances of receiving a favorable decision (we do the latter by assuming that the transition probabilities under manipulation first-order stochastically dominate those under improvement). We show how the firm's knowledge of the strategic behavior impacts its choice of a classifier, as well as agents' strategic choices and outcomes across different qualification states and demographic groups.

Specifically, we show that anticipating strategic behavior allows the firm to not only curb undesired manipulation (as also found in prior work) but can also incentivize agents to opt for improvement decisions instead. Notably, we find that when a group has high qualification rates, the firm will make its selection algorithm more "strict", incentivizing improvement by unqualified agents (driving them to improve both their qualification states and observable features), while still leaving the manipulation option open to qualified agents (who do not have sufficiently high observable features to be selected otherwise, but whose acceptance benefits the firm). We also highlight how a firm can leverage these strategic responses while enhancing algorithmic fairness.

II. Problem Setting and Preliminaries

We consider a Stackelberg game where the firm (algorithm designer) first announces a classifier, following which the agents respond strategically. We detail the agents' and the firm characteristics, their actions, and their utilities, in this section. The notation is summarized in Table I.

a) The agents: Consider a population of agents with two types of features: sensitive features that divide the population into two demographic groups $S \in \{a,b\}$ (e.g., race, gender), and an observable feature $x \in \mathbb{R}_{\geq 0}$ also used

Sura Alhanouti is with the Department of Integrated Systems Engineering, The Ohio State University. Parinaz Naghizadeh is with the Department of Electrical and Computer Engineering, University of California, San Diego. Emails: alhanouti.1@osu.edu, parinaz@ucsd.edu.

This work is supported by the NSF program on Fairness in AI in collaboration with Amazon under Award IIS-2040800. Any opinions, findings, and recommendations expressed in this material are those of the authors.

TABLE I: Summary of main notation.

Notation	Description
$s \in \{a,b\}$	Demographic group membership.
$n_s \in [0,1]$	Fraction of agents in group s.
$x \in \mathbb{R}_{\geq 0}$	Agents' observable feature pre-strategic behavior.
$y \in \{0,1\}$	Agents' unobservable, true qualification state pre-strategic behavior.
$\alpha_s \in [0,1]$	The qualification rate in group s pre-strategic behaviour.
$G_s^y(x)$	Feature distribution (pdf) for individuals with qualification y from group s pre-strategic behavior.
$\hat{x}, \hat{y}$, $\hat{\alpha}_s, \hat{G}_s^y(x)$	Agents' feature, qualification state, group qualification rate, and feature distribution, post-strategic behavior.
$w \in \{M,I,N\}$	Agents' strategic actions. Includes manipulation (M), improvement (I), and doing nothing (N).
$C_{w,s}$	Cost of strategic action w to an agent from group s.
$\tau_{w,s}^y(b)$, $\mathbb{T}_{w,s}^y(b)$	The pdf and CDF of the distribution of the boost b from action w for an agent from group s with true label y.
$\underline{b}_{w,s}^y / \bar{b}_{w,s}^y$	Minimum/maximum of boost distribution when taking action w for group s agents.
θ_s	The firm's decision threshold for group s.
u_+ / u_-	The firm's benefit/cost from true/false positives.
$\mathbb{I}_s^y / \mathbb{M}_s^y / \mathbb{N}_s^y$	Set of agents from group s with qualification y who opt for improvement/manipulation/doing nothing.

by the firm for making decisions (e.g., exam score, credit score). Let n_s be the fraction of agents in group s.

Each agent has a true hidden label or qualification state, denoted $Y \in \{0,1\}$, with $Y=1$ denoting a qualified agent and $Y=0$ denoting an unqualified agent. Let $\alpha_s := \mathbb{P}(Y=1|S=s)$ denote the qualification rate in group s. In addition, let $G_s^y(x) := \mathbb{P}(X=x|Y=y,S=s)$ denote the probability density function (pdf) of the distribution of the features for individuals with qualification state y from group s. We make the following assumption on the feature distributions.

Assumption 1: The feature distributions $G_s^y(x)$ are continuous, and satisfy the strict monotone likelihood ratio property: $\frac{G_s^1(x)}{G_s^0(x)}$ is strictly increasing in $x \in \mathbb{R}_{\geq 0}$.
In words, this entails that as an agent's feature increases, the likelihood that the agent is qualified increases as well.

b) The firm: A firm makes decisions on these agents based on their observable features x and their group memberships s. The decision is binary, denoted $d \in \{0,1\}$, where $d=1$ represents acceptance and $d=0$ represents rejection. This decision is determined by a group-dependent binary classifier $\pi_s(x) = \mathbb{P}(D=1|X=x,S=s)$. We assume that this is a threshold policy, such that if $x \geq \theta_s$ the probability to be accepted is $\pi_s(x)=1$, and is zero otherwise.[1]

c) Agents' strategic actions: After the policy is announced by the firm, agents have the option to behave strategically to improve their chances of receiving favorable outcomes. This is done by choosing one of the strategic actions $w \in \{M,I,N\}$, with M denoting *manipulation*, I denoting *improvement*, and N denoting *doing nothing*.

Taking these actions may impact the agents' features and/or labels. We use X and Y to denote the *pre-strategic* feature and label (before an action w is taken), and $\hat{X}$ and $\hat{Y}$ to denote the post-strategic ones. In particular, both qualified and unqualified agents ($y=1$ and $y=0$) can opt to manipulate ($w=M$) by changing their feature, $X=x \to \hat{X}=\hat{x}$, where $\hat{x}$ is some new feature, while the true label remains the same $Y=\hat{Y}=y$. Alternatively, both types of agents can choose to improve ($w=I$) by changing their feature $x \to \hat{x}$, as well as the true label $Y=y \to \hat{Y}=1$. Agents who opt for $w=N$ maintain their feature $\hat{X}=X=x$ and label $\hat{Y}=Y=y$. The firm will observe the agents' post-strategic features $\hat{x}$ when making its decision. We let $\hat{\alpha}_s$ and $\hat{G}_s^y(x)$ denote the *post-strategic* population statistics after these changes have happened as a result of agents' selected strategic actions.

In addition to differing in their impacts on changes in features and/or labels, the actions differ in two aspects: their cost and their efficacy. First, each action requires exerting effort and comes at a certain group-dependent (constant) cost $C_{w,s} \in [0,1)$, $w \in \{M,I,N\}, s \in \{a,b\}$. In addition, we assume all actions lead to a (weak) increase in the feature (i.e., we assume that $\hat{x} \geq x$), but that this increase in feature is different across actions: the probability that the feature x of a qualified/unqualified individual from a group s increases by b_w after opting for action w is distributed according to a *boost distribution* with pdf $\tau_{w,s}^y(b) := \mathbb{P}(\hat{X}=x+b|X=x,Y=y,W=w,S=s)$. These boost distributions determine the *efficacy* of the action. Let $\{\underline{b}_{w,s}^y, \bar{b}_{w,s}^y\}$ denote the minimum and maximum boost under action w, and $\mathbb{T}_{w,s}^y(b) : [\underline{b}_{w,s}^y, \bar{b}_{w,s}^y] \to [0,1]$ denote the CDF of the boost function.

We let $C_{N,s}=0$ and $\tau_{N,s}^y(0)=1$, meaning that the "do nothing" action has zero cost and no impact on changing the agent feature. The two remaining actions, M and I, can differ in cost and efficacy. We conduct our analysis under the following assumption.[2]

Assumption 2: Improvement is more costly than manipulation (i.e., $C_{I,s} \geq C_{M,s}$), and the improvement boost distribution first-order stochastically dominates (FOSD) that of manipulation (i.e., $\mathbb{T}_{M,s}^y(b) \geq \mathbb{T}_{I,s}^y(b)$).

In words, an action dominating another means that it "gets the agent further", in that it has a higher probability of increasing the agent's feature from x to a feature greater or equal to $\hat{x}=x+b$. Assumption 2 gives rise to the conflict between the two actions: manipulation is cheaper, while improvement is more effective in advancing the agent.

d) Firm's utility: The firm's goal is to find the optimal policy that maximizes its expected utility by correctly classifying the agents. The firm receives benefit u_+ from accepting qualified individuals, and incurs penalty u_- from accepting unqualified individuals. The firm's goal may alternatively be finding the *fair* optimal policy by also imposing a fairness constraint $\mathscr{C}$ on its decision problem.

Formally, let $U(\theta_a,\theta_b)$ represent the firm's total utility given the decision thresholds; the firm's expected utility is:

$$\mathbb{E}[U(\theta_a,\theta_b)] = \sum_s n_s \int_{\theta_s} [G_s^1(x)\alpha_s u_+ - G_s^0(x)(1-\alpha_s)u_-]\mathrm{d}x \ . \tag{1}$$

The firm's strategic (fair) optimization problem is to

[1] [13] shows that threshold policies are optimal under mild assumptions.

[2] There are three other scenarios: one can be recovered by interchanging the indexes M and I. The other two are perhaps less interesting, as one of the actions is more beneficial in both cost and efficacy in those cases.

choose the decision thresholds θ_a and θ_b as follows:

$$\max_{\theta_a,\theta_b} \quad \mathbb{E}[U(\theta_a,\theta_b)] \ , \qquad \textbf{s.t.} \quad \mathscr{C}_a^f(\theta_a) = \mathscr{C}_b^f(\theta_b). \tag{2}$$

The fairness constraint here, donated by f, can be, e.g., Equality of Opportunity ($\mathscr{C}_s^{EOP} = \int_{x\geq\theta_s} G_s^1(x)\mathrm{d}x$) or Demographic Parity ($\mathscr{C}_s^{DP} = \int_{x\geq\theta_s}(G_s^1(x)\alpha_s + G_s^0(x)(1-\alpha_s))\mathrm{d}x$).

When agents are strategic and this is known to the firm, the agents' statistics in (2) will be replaced by the post-strategic values, as characterized shortly in Section III.

e) Agents' utility: In general, each agent's strategic choice among the three available actions depends on its budget B, the cost of effort $C_{w,s}$, the efficacy of the selected action τ_{ws}^y, and the firm's deployed policy θ_s. For simplicity, we assume the same budget B for all agents, and assume this budget is sufficiently high so that all agents can afford both of the costly actions. Then, the choice among the actions depends on the relative benefit vs. the cost of each action.

Formally, an agent chooses to incur the cost $C_{w,s}$ of action w if and only if it increases the probability that the agent is accepted by the firm. For an agent from group s with feature x and qualification y, the benefit from strategic action w is

$$\begin{aligned}\mathscr{B}_{w,s}(x,y) := \mathbb{P}(D=1|X=x,Y=y,W=w,S=s) \\ - \mathbb{P}(D=1|X=x,Y=y,W=N,S=s).\end{aligned} \tag{3}$$

Note that the efficacy of the selected action $\tau_{w,s}^y$, and the firm's deployed policy θ_s, affect this benefit. The agent's utility will be $u_s(x,y,w) := \mathscr{B}_{w,s}(x,y) - C_{w,s}$. Note that the utility of action N is zero, capturing agents' outside option.

III. Agents' Strategic Behavior

We begin by characterizing agents' best responses to a given classifier. Consider an agent from group s with feature x and label y, facing decision threshold θ_s. Denote the agent's best-response by $w_s^*(x,y) := \arg\max_{w\in\{M,I,N\}} u_s(x,y,w)$. We will show that for any given y and s, the feature space x can be partitioned into disjoint regions determining the agents' best-reponse, with the boundaries of these regions determined by the points where agents become indifferent between pairs of actions. Specifically, we define a set of *indifference features* based on the cost-efficacy trade-off of the actions available to the agents.

Definition 1: Given efficacy CDFs $\mathbb{T}_{w,s}^y$, with inverse CDFs $(\mathbb{T}_{w,s}^y)^{-1}$, and costs $C_{w,s}$, define:

- **Opt in features:** $\mathbf{o}_{w,s}^y = \max\{0, \theta_s - (\mathbb{T}_{w,s}^y)^{-1}(1-C_{w,s})\}$, for $w \in \{M,I\}$.
- **Flip decision feature:** $\mathbf{f}_s^y \in [\theta - \bar{b}_M, \theta - \underline{b}_I]$ satisfying $\mathbb{T}_{M,s}^y(\theta_s - \mathbf{f}_s^y) - \mathbb{T}_{I,s}^y(\theta_s - \mathbf{f}_s^y) = C_{I,s} - C_{M,s}$, if a solution exists; zero otherwise.
- **Risk taker feature:** $\mathbf{r}_s^y := \max\{0, \theta_s - (\mathbb{T}_{M,s}^y)^{-1}(C_{I,s} - C_{M,s})\}$.

Intuitively, as shown formally in the proof of Proposition 1, these features can be interpreted as follows: The *opt in features* determine the first feature at which the agents benefit from opting for an action M or I as opposed to doing nothing; the *flip decision feature* is the feature at which the cost-efficacy trade-off of the M and I actions are equalized so that the agent will flip between its decisions around this point; and the risk taker feature is the feature at which the agent opts for an uncertain admission under M over a certain admission under I given M's relatively lower cost.

With the above, we are ready to characterize the agents' optimal best responses to a given decision threshold θ_s.

Proposition 1: Under Assumption 2, if $\mathbf{f}_s^y$ is unique, the agents' optimal response $w_s^*(x,y)$ to a given decision threshold θ_s will be one of the three types outlined in Table II.

TABLE II: Agents' best responses (Proposition 1).

Type	Condition	Range : $w_s^*(x,y)$
Type 1	$\mathbf{f}_s^y \leq \mathbf{o}_{I,s}^y \leq \mathbf{o}_{M,s}^y \leq \mathbf{r}_s^y$ *	$[0,\mathbf{o}_{I,s}^y) : N$, $[\mathbf{o}_{I,s}^y, \mathbf{r}_s^y) : I$, $[\mathbf{r}_s^y, \theta_s) : M$, $[\theta_s,\infty) : N$ †
Type 2	$\mathbf{o}_{M,s}^y \leq \mathbf{o}_{I,s}^y \leq \mathbf{f}_s^y$	$[0,\mathbf{o}_{M,s}^y) : N$, $[\mathbf{o}_{M,s}^y, \mathbf{f}_s^y) : M$, $[\mathbf{f}_s^y, \mathbf{r}_s^y) : I$, $[\mathbf{r}_s^y, \theta_s) : M$, $[\theta_s,\infty) : N$
Type 3	$\mathbf{f}_s^y \leq \mathbf{o}_{M,s}^y \leq \mathbf{o}_{I,s}^y$	$[0,\mathbf{o}_{M,s}^y) : N$, $[\mathbf{o}_{M,s}^y, \theta_s) : M$, $[\theta_s,\infty) : N$

* Or $\mathbf{o}_{I,s}^y \leq \mathbf{f}_s^y \leq \mathbf{o}_{M,s}^y \leq \mathbf{r}_s^y$.
† $[0,\mathbf{o}_{I,s}^y) : N$, $[\mathbf{o}_{I,s}^y, \mathbf{f}_s^y) : I$, $[\mathbf{f}_s^y, \theta_s) : M$, $[\theta_s,\infty) : N$.

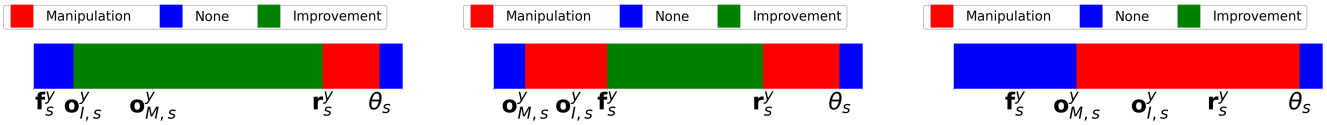
These three types of best-responses are illustrated in Figure 1. In particular, note that in all types of equilibrium, agents who are close to the decision threshold opt to be risk takers, choosing uncertain but cheap manipulation over certain but costly improvement. Interestingly, we posit that this may be consistent with gaming behavior observed in education settings: students committing academic dishonesty typically have higher GPAs [14].

We also note that the type of equilibrium is determined solely by the relative cost and efficacy of the manipulation and improvement actions; the choice of the classifier θ_s changes the indifference points where agents opt for each action, but not the *type* of equilibrium. We provide additional intuition for this below.

A. Illustration: uniform boost distributions

To further illustrate the intuition behind the types of best-responses identified in Proposition 1, we consider uniform boost distributions $\tau_{w,s}^y$. It is straightforward to verify that the three possible equilibria of Proposition 1 can be obtained by varying improvement cost C_I relative to manipulation cost C_M. Specifically, consider $\bar{b}_M^y - \underline{b}_M^y \geq \bar{b}_I^y - \underline{b}_I^y$. Then:

- **Low improvement cost**: If $C_M \leq C_I \leq \frac{\bar{b}_M^y - \underline{b}_M^y}{\bar{b}_I^y - \underline{b}_I^y} C_M + \frac{\bar{b}_I^y - \bar{b}_M^y}{\bar{b}_I^y - \underline{b}_I^y}$, the best-response is of Type 1 in Proposition 1. In this case, the improvement cost is relatively low, so coupled with its higher efficacy, agents find it beneficial to opt for improvement before manipulation becomes beneficial. Ultimately, however, the lower manipulation cost leads agents to change their decision closer to the threshold, once uncertainties about receiving a positive outcome are low enough.
- **Moderate improvement cost**: If $\frac{\bar{b}_M^y - \underline{b}_M^y}{\bar{b}_I^y - \underline{b}_I^y} C_M + \frac{\bar{b}_I^y - \bar{b}_M^y}{\bar{b}_I^y - \underline{b}_I^y} \leq C_I \leq C_M + \frac{\underline{b}_I^y - \underline{b}_M^y}{\bar{b}_M^y - \underline{b}_M^y}$, the best-response is of Type 2 in Proposition 1. Here, improvement costs are too high to benefit the agents who are far from the decision thresh-



(a) Type 1. (b) Type 2. (c) Type 3.

Fig. 1: Agent best-responses identified in Proposition 1. The axis shows agents' features x, and colors show their action.

old, but low enough so that agents opt for improvement over manipulation when both actions are uncertain.

- **High improvement cost**: If $C_M + \frac{b_I^y - b_M^y}{b_M^y - b_M^y} \leq C_I$, the best-response is of Type 3 in Proposition 1. In this case, the improvement cost is significantly higher than the manipulation cost, leading all agents to pick manipulation when profitable.

B. Post-strategic population statistics

We next characterize the *post-strategic* population statistics. In particular, as noted earlier, the strategic actions $w \in \{M, I\}$ change the agent's true qualification and/or feature. Denote the post-strategic qualification rates by $\hat{\alpha}_s = \mathbb{P}(\hat{Y} = 1 | S = s)$, and the post-strategic feature distribution by $\hat{G}_s^y(x) := \mathbb{P}(\hat{X} = x | \hat{Y} = y, S = s)$. Let $\mathbb{I}_s^y := \{x | Y = y, S = s, w = I\}$ be the set of agents from group s with label y who opt for improvement under a given threshold policy θ_s. Define $\mathbb{M}_s^y$ and $\mathbb{N}_s^y$ similarly for the set of agents who opt for manipulation and for doing nothing. These sets of agents are characterized by the best responses of Proposition 1 under different problem parameter settings.

The following lemma presents the expressions for the post-strategic population statistics in terms of the pre-strategic population statistics and the best-response regions.

Lemma 1: The post-strategic qualification rate is given by

$$\hat{\alpha}_s = \alpha_s + (1 - \alpha_s) \int_{x \in \mathbb{I}_s^0} G_s^0(x) \mathrm{d}x \tag{4}$$

The post-strategic feature distributions are given by:

$$\hat{G}_s^0(x) = \frac{1 - \alpha_s}{1 - \hat{\alpha}_s} \Big(\mathbb{1}(x \in \mathbb{N}_s^0) G_s^0(x) + (G_s^0 * \tau_{M,s}^0)(x) \Big). \tag{5}$$

and,

$$\hat{G}_s^1(x) = \frac{\alpha_s}{\hat{\alpha}_s} \Big(\mathbb{1}(x \in \mathbb{N}_s^1) G_s^1(x) + (G_s^1 * \tau_{M,s}^1)(x) + (G_s^1 * \tau_{I,s}^1)(x) \Big) + \frac{1 - \alpha_s}{\hat{\alpha}_s} (G_s^0 * \tau_{I,s}^0)(x), \tag{6}$$

where $(G_s^y * \tau_{w,s}^y)(x) := \int_{z \in \mathbb{W}_s^y} G_s^y(z) \tau_{w,s}^y(x - z) \mathrm{d}z$ is the convolution of the feature distribution and boost function restricted to the action w region.

IV. Firm's Optimal Policy

Using the findings of Section III, we can now determine the firm's optimal choice of decision thresholds.

We start with a firm who does not account for agents' strategic behavior and does not implement any fairness constraints; we refer to this as the *unfair non-strategic* firm. The lemma below characterizes such firm's optimal decision thresholds as a function of the population statistics.

Lemma 2: The unfair non-strategic firm's optimal decision thresholds θ_s^U satisfies $\frac{G_s^1(\theta_s^U)}{G_s^0(\theta_s^U)} = \frac{u_-(1-\alpha_s)}{u_+ \alpha_s}$.

This is similar to results obtained in prior work [15], [16]. (The proof is omitted in interest of space.)

If the firm is cognizant of the agents' strategic behavior, the *unfair strategic* firm's optimal thresholds $\hat{\theta}_s^U$ can be obtained by finding the utility maximizer when the firm's utility (1) is evaluated on the population post-strategic statistics characterized in Lemma 1. Specifically, the strategic firm's utility can be expressed as follows:

$$\begin{aligned} \hat{U}(\theta_a^U, \theta_b^U) := U(\theta_a^U, \theta_b^U) \\ + \sum_s u_+ \alpha_s \Phi_{s,(i)}^1(\theta_s) + u_+ (1 - \alpha_s) \Phi_{s,(i)}^0(\theta_s) \\ + u_+ \alpha_s \Psi_{s,(i)}^1(\theta_s) - u_- (1 - \alpha_s) \Psi_{s,(i)}^0(\theta_s), \end{aligned} \tag{7}$$

where *(i)* denotes the type of best response (as identified in Proposition 1), and

$$\begin{aligned} \Phi_{s,(i)}^y(\theta_s) &= \int_{z \in \mathbb{I}_s^y} \big(G_s^y(z) - G_s^y(z) \mathbb{T}_I^y(\theta_s - z) \big) \mathrm{d}z, \\ \Psi_{s,(i)}^y(\theta_s) &= \int_{z \in \mathbb{M}_s^y} \big(G_s^y(z) - G_s^y(z) \mathbb{T}_M^y(\theta_s - z) \big) \mathrm{d}z. \end{aligned} \tag{8}$$

Intuitively, the term $\Phi_{s,(i)}^y(\theta_s)$ (resp. $\Psi_{s,(i)}^y(\theta_s)$) captures the change in the firm's utility due to label y agents from group s who opt for improvement (resp. manipulation) when facing classifier θ_s. Note that the first three of these terms *increase* the firm's utility over the non-strategic utility: these are all agents who improved, and the qualified agents who manipulate. The last term, with the negative sign, decreases the firm's utility, as these are unqualified agents who pass the threshold θ_s through manipulation. Accordingly, we can characterize the unfair strategic firm's optimal policy.

Lemma 3: The unfair strategic firm's optimal decision thresholds $\hat{\theta}_s^U$ for best-response type (i) satisfy $\frac{\Phi_{s,(i)}^{\prime 1} + \Psi_{s,(i)}^{\prime 1} - G_s^1(\hat{\theta}_s^U) + \frac{(1-\alpha_s)}{\alpha_s} \Phi_{s,(i)}^{\prime 0}}{\Psi_{s,(i)}^{\prime 0} - G_s^0(\hat{\theta}_s^U)} = \frac{u_-(1-\alpha_s)}{u_+ \alpha_s}$, where $\Phi_{s,(i)}^{\prime y}$ ($\Psi_{s,(i)}^{\prime y}$) is the first order derivative of $\Phi_{s,(i)}^y$ ($\Psi_{s,(i)}^y$) w.r.t $\hat{\theta}_s$.

The detailed derivation of the thresholds under each best-response type can be found in Table III. The main challenge in finding these optimal thresholds is that the regions $\mathbb{I}_s^y$ and $\mathbb{M}_s^y$ in which agents opt for the improvement and manipulation actions are functions of the decision variables θ_s, and these dependencies should be accounted for when evaluating the derivatives $\Phi_{s,(i)}^{\prime y}$ and $\Psi_{s,(i)}^{\prime y}$. This is done by applying the Leibniz integral rule, and then leveraging the relation between the threshold and the indifference points in agents' strategic responses. (The detailed derivations are omitted in interest of space.)

Intuitively, the terms in Table III can be interpreted as follows. Take $(1-\alpha_s)\Big(C_{I,s}G_s^0(\mathbf{o}_{I,s}^0)+(G_s^0 * \tau_{I,s}^0)(\hat{\theta}_s^U)-G_s^0(\mathbf{r}_s^0)\Big)$, appearing in the numerator of Type 1 equilibrium characterization, which reflects the rate of change in the benefits from accepting unqualified agents who opt for improvement, as the decision threshold changes. If the decision threshold $\hat{\theta}_s^U$ increases by a small ε, the firm will lose the agents at $x=\mathbf{o}_{I,s}^0$ who successfully made it to the old threshold through improvement (at a rate $(1-\alpha_s)C_{I,s}G_s^0(\mathbf{o}_{I,s}^0)$), will lose some of those with $x\in(\mathbf{o}_{I,s}^0,\mathbf{r}_{I,s}^0)$ who no longer make it to the new threshold (at a rate $(1-\alpha_s)(G_s^0 * \tau_{I,s}^0)(\hat{\theta}_s^U)$), yet will *gain* from agents with $x=\mathbf{r}_s^0$ who now opt for improvement instead of manipulation (at a rate $(1-\alpha_s)G_s^0(\mathbf{r}_s^0)$). Other expressions can be interpreted similarly.

Lastly, we extend the above two lemmas when the firm also incorporates a fairness constraint in its selection.

Lemma 4: The fair non-strategic firm's optimal decision thresholds $\theta_s^{\mathscr{C}}$ satisfy: $\sum_s n_s \frac{u_+\alpha_s G_s^1(\theta_s^{\mathscr{C}})-u_-(1-\alpha_s)G_s^0(\theta_s^{\mathscr{C}})}{\frac{\partial \mathscr{C}_s^f(\theta_s^{\mathscr{C}})}{\partial \theta_s^{\mathscr{C}}}}=0$.

Lemma 5: The fair strategic firm's optimal decision thresholds $\hat{\theta}_s^{\mathscr{C}}$ satisfy: $\sum_s n_s \frac{u_+\alpha_s(\Phi'^1_{s,(i)}+\Psi'^1_{s,(i)}-G_s^1(\hat{\theta}_s^{\mathscr{C}}))+(1-\alpha_s)(u_+\Phi'^0_{s,(i)}-u_-(\Psi'^0_{s,(i)}-G_s^0(\hat{\theta}_s^{\mathscr{C}})))}{\frac{\partial \mathscr{C}_s^f(\hat{\theta}_s^{\mathscr{C}})}{\partial \theta_s^{\mathscr{C}}}}=0$.

V. Effects of Predicting Strategic Behavior

We can now proceed to analyzing the impacts of anticipating agents' strategic behavior on the optimal policies and the firm's utility by comparing the strategic policy $\hat{\theta}_s$ (from Lemma 3) with the non-strategic policy θ_s (from Lemma 2). We do so under Types 1 and 3 of agents' best responses, which can be interpreted as relatively low and high improvement costs, respectively, as illustrated in Section III-A. Notably, Type 3 equilibria only include manipulation decisions, whereas Type 1 includes both manipulation and improvement decisions at equilibrium. Contrasting these types of equilibria allows us to compare our findings with prior work, e.g. [15], which have studied the impact of anticipating gaming given only manipulation decisions.

To do so, first note that when a decision threshold θ_s is lowered (resp. increased), more (resp. fewer) agents are accepted without taking any strategic action. Motivated by this, we begin with the following definition, which specifies decision thresholds at which improvement/manipulation decisions have their highest and lowest possible impact.

Definition 2: For Type *(i)* best-responses, define:

- $\overline{\Phi}^y_{s,(i)} := \{\hat{\theta}_s \in \arg\max_{\theta_s} \Phi^y_{s,(i)}(\theta_s)\}$; similarly for $\overline{\Psi}^y_{s,(i)}$.
- $\underline{\Phi}^y_{s,(i)} := \{\hat{\theta}_s \in \arg\min_{\theta_s} \Phi^y_{s,(i)}(\theta_s)\}$; similarly for $\underline{\Psi}^y_{s,(i)}$.

Recall that $\Phi^y_{s,(i)}(\theta_s)$ captures the change in the firm's utility due to label y agents from group s who opt for improvement. Then, in words, $\overline{\Phi}^y_{s,(i)}$(resp. $\underline{\Phi}^y_{s,(i)}$), which is defined for Type 1 equilibria, denotes the value(s) of θ_s at which label y agents from group s opting for improvement would have their *maximum* (resp. *minimum*) impact on the firm's utility. The minimum possible impact is zero, which a threshold in $\underline{\Phi}^y_{s,(i)}$ can lead to for one of two reasons: either θ_s is so high that none of the label y agents can advance to θ_s even when opting for improvement, or so low that all label y agents can be accepted with their original feature x. (Formally, in these cases, the corresponding indifference features defined in Definition 1 are outside the range of agents' feature distributions). On the other hand, $\overline{\Phi}^y_{s,(i)}$ are the θ_s under which the firm can induce the highest proportion of label y agents in group s to succeed at improvement. The terms $\overline{\Psi}^y_{s,(i)}$ and $\underline{\Psi}^y_{s,(i)}$, which are defined in both Type 1 and 3 equilibria, can be interpreted similarly.

Table IV outlines where the optimal policies lie relative to these extreme values,[3] under low vs. high base qualification rates. Specifically, let $\xi=\frac{u_-}{u_-+u_+}$. Then, $\alpha_s>\xi$ (resp. $\alpha_s<\xi$) means that from the viewpoint of the firm, when accounting for the relative costs of true vs. false positives, most of the agents in group s are qualified (resp. unqualified) before any strategic actions are taken. We refer to this as a *majority qualified* (resp. *majority unqualified*) group.[4]

a) Majority-unqualified (low α_s); Type 3 response: Here, a nonstrategic firm choosing θ_s focuses on rejecting the many unqualified agents by choosing a relatively high θ_s. Strategic agents can still opt for manipulation to increase their chances of surpassing the threshold. A strategic firm accounts for this feedback loop, and increases $\hat{\theta}_s$ even further (this is seen in the top right quadrant of Table IV). The higher threshold accommodates less manipulation by unqualified ($Y=0$) agents who are now too far from the threshold to benefit from manipulation, but leads to more manipulation by qualified ($Y=1$) agents who are no longer accepted by default. Intuitively, this is the firm's desired strategic response by agents, as the group is majority unqualified, and manipulation by qualified agents *benefits* the firm.

b) Majority-unqualified (low α_s); Type 1 response: In a Type 1 equilibrium, agents also have opportunities to opt for improvement decisions; this is however not a consideration in a non-strategic firm's decision. Such firm chooses a relatively high threshold θ_s, only with the intent of rejecting more of the majority-unqualified group. A strategic firm, on the other hand, understands that further increasing the threshold $\hat{\theta}_s$ relative to θ_s not only reduces manipulation opportunities, but also increases the benefits from agents who opt for improvement. Formally, increasing $\hat{\theta}_s$ also increases the indifference points $\mathbf{r}_s^y$ (the feature at which agents flip their decision from improvement to manipulation) and $\mathbf{o}_{I,s}^y$ (the feature at which agents first benefit from improvement). The former increase in $\mathbf{r}_s^y$ benefits the firm as there are more qualified ($Y=1$) and less unqualified ($Y=0$) agents in the new manipulation range (between the new $\mathbf{r}_s^y$ and $\hat{\theta}_s$); this is the same observation made in Type 3 equilibria above, and is in line with effects noted in prior work (e.g., [15]). The latter increase in $\mathbf{o}_{I,s}^y$ also benefits the firm, as

[3] As an example, the inequality signs in the table reflect the following: $\theta_s>\overline{\Phi}^y_{s,(i)}$ implies the selected threshold is not in (and specifically, greater than) the values which would lead to the maximum proportion of label y agents from group s to strategically select action I.

[4] In some real-world data, e.g., FICO credit scores [17], advantaged (resp. disadvantaged) groups are found to be majority-qualified (resp. unqualified).

TABLE III: Optimal decision thresholds in Lemma 3.

Type	**Type 1**	**Type 2**	**Type 3**
Numerator	$C_{I,s}G_s^1(\mathbf{o}_{I,s}^1) + (G_s^1 * \tau_{I,s}^1)(\hat{\theta}_s^U) - (C_{I,s} - C_{M,s})G_s^1(\mathbf{r}_s^1) + (G_s^1 * \tau_{M,s}^1)(\hat{\theta}_s^U) + \frac{(1-\alpha_s)}{\alpha_s}\Big(C_{I,s}G_s^0(\mathbf{o}_{I,s}^0) + (G_s^0 * \tau_{I,s}^0)(\hat{\theta}_s^U) - G_s^0(\mathbf{r}_s^0)\Big)$	$C_{M,s}G_s^1(\mathbf{o}_{M,s}^1) + (G_s^1 * \tau_{I,s}^1)(\hat{\theta}_s^U) + (G_s^1 * \tau_{M,s}^1)(\hat{\theta}_s^U) + G_s^1(\mathbf{f}_s^1)(C_{I,s} - C_{M,s}) - G_s^1(\mathbf{r}_s^1)(C_{I,s} - C_{M,s}) + \frac{(1-\alpha_s)}{\alpha_s}\Big(G_s^0(\mathbf{f}_s^0)(1 - \mathbb{T}_{I,s}^0(\hat{\theta}_s^U - \mathbf{f}_s^0)) - G_s^0(\mathbf{r}_s^0) + (G_s^0 * \tau_{I,s}^0)(\hat{\theta}_s^U)\Big)$	$C_{M,s}G_s^1(\mathbf{o}_{M,s}^1) + (G_s^1 * \tau_{M,s}^1)(\hat{\theta}_s^U)$
Denominator	$(1-(C_{I,s}-C_{M,s}))G_s^0(\mathbf{r}_s^0) + (G_s^0 * \tau_{M,s}^0)(\hat{\theta}_s^U)$	$C_{M,s}G_s^0(\mathbf{o}_{M,s}^0) - (1-\mathbb{T}_{M,s}^0(\hat{\theta}_s^U - \mathbf{f}_s^0))G_s^0(\mathbf{f}_s^0) + (1-(C_{I,s}-C_{M,s}))G_s^0(\mathbf{r}_s^0) + (G_s^0 * \tau_{M,s}^0)(\hat{\theta}_s^U)$	$C_{M,s}G_s^0(\mathbf{o}_{M,s}^0) + (G_s^0 * \tau_{M,s}^0)(\hat{\theta}_s^U)$

TABLE IV: Strategic vs non-strategic optimal thresholds.

α	$(0,\xi)$		$(\xi,1)$	
	Strategic	Non-Str.	Strategic	Non-Str.
Type 1	$\underset{\overline{\Phi}_{(1)}^1}{\in} \in \overline{\Phi}_{(1)}^0$	$\underset{\overline{\Phi}_{(1)}^1}{\leq} \underset{\overline{\Phi}_{(1)}^0}{\leq}$	$\underset{\underline{\Phi}_{(1)}^1}{\geq} \underset{\overline{\Phi}_{(1)}^0}{\leq}$	$\underset{\underline{\Phi}_{(1)}^1}{\in} \underset{\overline{\Phi}_{(1)}^0}{\leq}$
	$\underset{\overline{\Psi}_{(1)}^1}{\in} \leq \underline{\Psi}_{(1)}^0$	$\underset{\overline{\Psi}_{(1)}^1}{\in} \underset{\overline{\Psi}_{(1)}^0}{\in}$	$\underset{\overline{\Psi}_{(1)}^1}{\in} \underset{\overline{\Psi}_{(1)}^0}{\leq}$	$\underset{\underline{\Psi}_{(1)}^1}{\in} \underset{\overline{\Psi}_{(1)}^0}{\in}$
Type 3	$\underset{\overline{\Psi}_{(3)}^1}{\geq} (\overline{\Psi}_{(3)}^0, \underline{\Psi}_{(3)}^0]$	$\underset{\overline{\Psi}_{(3)}^1}{\leq} \underset{\overline{\Psi}_{(3)}^0}{\in}$	$\underset{\overline{\Psi}_{(3)}^1}{\leq} \underset{\overline{\Psi}_{(3)}^0}{\leq}$	$\underset{\underline{\Psi}_{(3)}^1}{\in} \underset{\overline{\Psi}_{(3)}^0}{\leq}$

there are now more agents (in both label $Y=0$ and $Y=1$) opting for improvement who successfully improve their true qualification states, are accepted by the firm, and positively impact its utility. This effect is new to our model, and highlights how anticipating gaming can lead to increased improvement incentives among agents. These effects are illustrated in the top left quadrant of Table IV.

c) Majority-qualified (high α_s); Type 3 response: Here, we observe similar effects as before, in that a strategic firm chooses a higher threshold than a non-strategic firm to control strategic manipulation (as illustrated in the bottom right quadrant of Table IV). Interestingly, the strategic firm might overall *increase* the number of agents who get accepted through manipulation, as the increase in threshold would drive the many qualified ($Y=1$) agents who are no longer accepted by default to choose M (note however that the change still decreases the number of unqualified agents ($Y=0$) who can pass the threshold through manipulation).

d) Majority-qualified (high α_s); Type 1 response: For Type 1 best-responses, the impacts of anticipating strategic behavior on the firm's utility is more significant in the majority-qualified case compared to the majority-unqualified case. When the majority are qualified, the strategic policy can select a threshold that still allows qualified agents to manipulate, keeping similar or fewer manipulation opportunities for unqualified agents. This threshold can also motivate more qualified and unqualified agents to improve (due to similar reasons noted in the majority-unqualified case). These effects are illustrated in the bottom left quadrant of Table IV.

VI. Numerical Illustration

We further illustrate the findings of Section V through a numerical example. We consider a population consisting of two equal size groups a and b, with group a being majority-qualified ($\alpha_a = 0.7$) and group b being majority-unqualified ($\alpha_b = 0.2$). The feature distributions for both groups follows Gaussian distributions, with $G_a^0 = N(70,15^2), G_b^0 = N(60,15^2), G_a^1 = N(110,15^2)$, and $G_b^1 = N(90,15^2)$. We let the boost distributions follow uniform distributions, with $\tau_{I,s}^1(b) = Uniform(40,80), \tau_{I,s}^0(b) = Uniform(35,77)$, and $\tau_{M,s}^y(b) = Uniform(10,75)$, and let the costs of the decisions be $C_{I,s} = 0.3$, and $C_{M,s} = 0.2$, so that the agents' best-responses follow a Type 1 (manipulation and improvement) equilibrium. Note that we assume the same costs across groups $s \in \{a,b\}$, with disparities arising due to the difference in their feature distributions.

We compare the non-strategic and strategic (Table V) firms' utilities and thresholds, and their impact on how agents from different groups and labels opt for manipulation vs. improvement decisions. Specifically, the rows I_s^y and M_s^y show the percentage of pre-strategic label y agents from group s who choose the respective action (out of the total population of agents), and $\%_{w,s}^y$ denotes the percentage of post-strategic label y agents from group s opting for action w who successfully get accepted.

First, we note that the observations made in Section V are reflected in these experiments: for both the majority-qualified group a and the majority-unqualified group b, a strategic firm incentivizes improvement decisions. It also discourages manipulation in both groups, but discourages it significantly more in the majority-unqualified group b (while allowing for *more* manipulation by *qualified* agents in both groups).

Next, we contrast the pre-strategic α_s and post-strategic $\hat{\alpha}_s$ qualification rates for each group. First, whether the firm is strategic or not, these rates increase regardless of the use of fairness constraints and across groups. This is due to the availability of the improvement option. We also note that the strategic firm is more successful at incentivizing improvement actions, as evidenced by the higher $\hat{\alpha}_s$ in both groups a and b compared to the non-strategic firm. More interestingly, the majority-unqualified group b becomes majority-qualified in both fair and unfair policies implemented by the strategic firm. This is however not the case with DP-Fair and TPR-Fair policies selected by a non-strategic firm. Such firm does not account for agents' strategic responses, lowering the decision threshold on group b and increasing it on group a, in order to satisfy the fairness constraint. This reduces the motivation of agents from group b to improve, as more are accepted by default, while also enabling other (majority-unqualified) agents from this group to pass the threshold through manipulation. In contrast, a non-strategic firm lowers its fair thresholds on group b much less drastically, as it realizes that parity between selection rates (as required by DP) and true-positive rates (as required by TPR) can be achieved by a combination of adjusting the thresholds *and* driving agents' best-responses.

TABLE V: Comparison of Non-strategic and Strategic Policies.

	Non-strategic policy								**Strategic policy**							
	UnFair		**DP-Fair**		**TPR-Fair**				**UnFair**		**DP-Fair**		**TPR-Fair**			
U_{Total}	48.32		27.85		41.72				81.17		74.97		69.49			
Group	a		b		a		b		a		b		a		b	
$U(\theta_s)$	64.23		33.72		71.27		-14.83		77.14		6.95		88.77		73.75	
θ_s	80		78		86.45		55.11		88.6		68.63		106.87		97.25	
α_s	0.7		0.2		0.7		0.2		0.7		0.2		0.7		0.2	
$\hat{\alpha}_s$	0.79		0.62		0.85		0.26		0.87		0.44		0.96		0.90	
$\hat{Y}$	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
I	0.1	9.9	0.5	42.9	0.2	15	0	6.1	0.4	16.7	0.1	24	6.7	26.5	5.3	70.5
M	1.6	12.6	3.5	27.9	3.7	11	0.2	23.7	4.9	10	1.4	33.3	22.6	2.5	8.3	6.1
$\%^y_{I,s}$	9.8	-	41.7	-	14.7	-	6.1	-	16.5	-	23.7	-	29	-	64.5	-
$\%^y_{M,s}$	1.6	12.3	3.5	27.2	3.7	10.7	0.1	23.4	4.9	9.7	1.4	32.7	22.2	2.4	8.1	5.8
Total	70.6	19.9	61	36.4	84.2	14.8	26	73.6	86	13	44.6	55.3	91.8	2.6	78.9	6.4

	UnFair		**DP-Fair**		**TPR-Fair**	
	a	b	a	b	a	b
$U(\theta_s)$	88.77	73.75	84.88	66.29	85.73	54.78
θ_s	106.87	97.25	116.68	90.305	116.9	84.815
α_s	0.7	0.2	0.7	0.2	0.7	0.2
$\hat{\alpha}_s$	0.96	0.90	0.96	0.85	0.96	0.76
$\hat{Y}$	1 0	1 0	1 0	1 0	1 0	1 0
I	18.1 25.9	2.8 64.8	18.3 25.6	1.4 56.5		
M	28.8 0.6	7.3 12.7	29.2 0.6	5.8 19.3		
$\%^y_{I,s}$	37.3 -	60.7 -	37.3 -	54 -		
$\%^y_{M,s}$	28.3 0.6	7.2 12.3	28.7 0.6	5.7 18.8		
Total	88.5 0.7	77.6 14.1	88.5 0.6	72.3 22.7		

significant decrease. decrease. increase. significant increase.

VII. Conclusion and Future Work

We proposed a Stackelberg game model to study strategic classification, where a firm deploys a (fair) classifier and agents strategically respond by adjusting their true qualification states and/or observable features to increase their chances of acceptance. We model both different costs and stochastic efficacy for the agents' manipulation and improvement actions. We find that anticipating strategic behavior can allow the firm to not only curb manipulation behavior (as also noted in prior work) but can also incentivize agents to opt for improvement decisions. Specifically, we find that a strategic firm chooses its policy to incentivize improvement by unqualified agents (driving them to improve both their qualification states and observable features), while still allowing for manipulation by some qualified agents (who do not have sufficiently high observable features to be selected otherwise). Our numerical experiment highlights that the firm can leverage agents' strategic behavior (mainly, improvement decisions) to satisfy fairness constraints without drastically adjusting its selection rule; analytical support for this observation is a main direction of our future work.

References

[1] M. Möhlmann and L. Zalmanson, "Hands on the wheel: Navigating algorithmic management and uber drivers'," in *Autonomy, in proceedings of the international conference on information systems*, 2017.

[2] "5 resume hacks to pass ATS," https://www.forbes.com/sites/ashleystahl/2022/12/12/5-resume-hacks-to-pass-ats/?sh=3668530d4b2b, accessed: 2024-03-15.

[3] M. Eslami, K. Karahalios, C. Sandvig, K. Vaccaro, A. Rickman, K. Hamilton, and A. Kirlik, "First i" like" it, then i hide it: Folk theories of social feeds," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016.

[4] S. Levanon and N. Rosenfeld, "Generalized strategic classification and the case of aligned incentives," in *International Conference on Machine Learning*, 2022.

[5] T. Lechner and R. Urner, "Learning losses for strategic classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7337–7344.

[6] S. Levanon and N. Rosenfeld, "Strategic classification made practical," in *International Conference on Machine Learning*, 2021.

[7] S. Milli, J. Miller, A. D. Dragan, and M. Hardt, "The social cost of strategic classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 230–239.

[8] Y. Shavit, B. Edelman, and B. Axelrod, "Causal strategic linear regression," in *International Conference on Machine Learning*, 2020.

[9] Y. Bechavod, K. Ligett, S. Wu, and J. Ziani, "Gaming helps! learning from strategic interactions in natural dynamics," in *International Conference on Artificial Intelligence and Statistics*, 2021.

[10] K. Jin, X. Zhang, M. M. Khalili, P. Naghizadeh, and M. Liu, "Incentive mechanisms for strategic classification and regression problems," in *The 23rd ACM Conference on Economics and Computation*, 2022.

[11] T. Alon, M. Dobson, A. Procaccia, I. Talgam-Cohen, and J. Tucker-Foltz, "Multiagent evaluation mechanisms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[12] K. Jin, Z. Huang, and M. Liu, "Collaboration as a mechanism for more robust strategic classification," in *62nd IEEE Conference on Decision and Control (CDC)*, 2023.

[13] X. Zhang, R. Tu, Y. Liu, M. Liu, H. Kjellstrom, K. Zhang, and C. Zhang, "How do fair decisions fare in long-term qualification?" *Advances in Neural Information Processing Systems*, 2020.

[14] "8 astonishing stats on academic cheating," https://www.oedb.org/ilibrarian/8-astonishing-stats-on-academic-cheating/, accessed: 2024-03-15.

[15] X. Zhang, M. M. Khalili, K. Jin, P. Naghizadeh, and M. Liu, "Fairness interventions as (dis) incentives for strategic manipulation," in *International Conference on Machine Learning*, 2022.

[16] Y. Liao and P. Naghizadeh, "Social bias meets data bias: The impacts of labeling and measurement errors on fairness criteria," *The Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023.

[17] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, 2016.

Appendix

A. Proof of Proposition 1

We begin by establishing the possible orderings between the indifference points of Definition 1, under Assumption 2, showing that the cases identified in Proposition 1 cover all possible equilibrium outcomes given the possible orderings of the four constants $\mathbf{o}^y_{M,s}, \mathbf{o}^y_{I,s}, \mathbf{f}^y_s, \mathbf{r}^y_s$.

Lemma 6: We have $\mathbf{r}^y_s \geq \{\mathbf{f}^y_s, \mathbf{o}^y_{M,s}\}$. Further, if $\mathbf{f}^y_s$ is unique, either $\mathbf{f}^y_s \geq \max\{\mathbf{o}^y_{M,s}, \mathbf{o}^y_{I,s}\}$, or $\mathbf{f}^y_s \leq \min\{\mathbf{o}^y_{M,s}, \mathbf{o}^y_{I,s}\}$. Additionally, if $\mathbf{o}^y_{I,s} \leq \mathbf{o}^y_{M,s}$, there is at most one $\mathbf{f}^y_s \leq \mathbf{o}^y_{I,s}$.

Proof: First, note that $(\mathbb{T}^y_{M,s})^{-1}(C_I - C_{M,s}) \leq (\mathbb{T}^y_{M,s})^{-1}(1 - C_{M,s})$; this is because the inverse CDF is an increasing function, and $C_{I,s} \leq 1$. Therefore, $\mathbf{o}^y_{M,s} \leq \mathbf{r}^y_s$. Also, by definition, $\mathbb{T}^y_{M,s}(\theta_s - \mathbf{f}^y_s) - \mathbb{T}^y_{I,s}(\theta_s - \mathbf{f}^y_s) = \mathbb{T}^y_{M,s}(\theta_s - \mathbf{r}^y_s)$, and therefore $\mathbf{f}^y_s \leq \mathbf{r}^y_s$. Finally, note that if $\mathbf{f}^y_s \geq \mathbf{o}^y_{M,s}$, then $\mathbb{T}^y_{I,s}(\theta_s - \mathbf{f}^y_s) = \mathbb{T}^y_{M,s}(\theta_s - \mathbf{f}^y_s) - (C_{I,s} - C_{M,s}) \leq \mathbb{T}^y_{M,s}(\theta_s - \mathbf{o}^y_{M,s}) - (C_{I,s} - C_{M,s}) = 1 - C_{I,s} = \mathbb{T}^y_{I,s}(\theta_s - \mathbf{o}^y_{I,s})$, and therefore $\mathbf{f}^y_s \geq \mathbf{o}^y_{I,s}$. Similarly, we can show that if $\mathbf{f}^y_s \leq \mathbf{o}^y_{M,s}$, then $\mathbf{f}^y_s \leq \mathbf{o}^y_{I,s}$. Lastly, if $\mathbf{o}^y_{I,s} \leq \mathbf{o}^y_{M,s}$, and given that $u_s(0,y,I) < u_s(0,y,M)$ due to FOSD, it must be that the utility of I crosses the utility of M at some $x \leq \mathbf{o}^y_{I,s}$, and therefore $\mathbf{f}^y_s \leq \mathbf{o}^y_{I,s}$. ■

Now, note that if $x \geq \theta_s$, the agent is already admitted by the classifier, and finds it optimal to do nothing. As

such, only agents with $x < \theta_s$ may opt for manipulation or improvement decisions. For these agents, the probability of being admitted if neither action is taken is zero. Together with (3), this means that the utility of such agents with $x < \theta_s$ when choosing $w \in \{M, I\}$ reduces to $u_s(x, y, w) = \mathbb{P}(\hat{x} \geq \theta_s | X = x, Y = y, W = w, S = s) - C_{w,s}$.

We now proceed by finding the features $\mathbf{o}^y_{M,s}$ at which the agent first finds it beneficial to opt for manipulation over doing nothing. Note that the agent's utility is non-decreasing in x. This is because we have assumed adopting either of the two actions M or I weakly increases the agent's feature, and hence (weakly) increases the probability of being admitted by the classifier. This means that if an agent with feature $\bar{x}$ prefers action w over doing nothing, so will all $x > \bar{x}$.

Recall that $u_s(x, y, N) = 0$. Therefore, $\mathbf{o}^y_{M,s}$ is the first x at which $u_s(x, y, M) \geq 0$. This is given by:

$$\begin{aligned} &\mathbb{P}(\hat{x} \geq \theta_s | X = x, Y = y, W = M, S = s) \geq C_{M,s} \\ \Leftrightarrow \quad &x \geq \theta_s - (\mathbb{T}^y_{M,s})^{-1}(1 - C_{M,s}) \ . \end{aligned}$$

Therefore, the first point at which the agent finds it beneficial to opt for M over N is $\mathbf{o}^y_{M,s} = \max\{0, \theta_s - (\mathbb{T}^y_{M,s})^{-1}(1 - C_{M,s})\}$. The first point at which the agent benefits from I over N can be similarly found to be $\mathbf{o}^y_{I,s} = \max\{0, \theta_s - (\mathbb{T}^y_{I,s})^{-1}(1 - C_{I,s})\}$. Let $z := \arg\min_{\{M,I\}}\{\mathbf{o}^y_{I,s}, \mathbf{o}^y_{M,s}\}$. Then, agents with $0 \leq x < \min\{\mathbf{o}^y_{I,s}, \mathbf{o}^y_{M,s}\}$ opt for N, while those with $\min\{\mathbf{o}^y_{I,s}, \mathbf{o}^y_{M,s}\} \leq x < \max\{\mathbf{o}^y_{I,s}, \mathbf{o}^y_{M,s}\}$ opt for action z.

Next, given that the improvement action first-order stochastically dominates the manipulation action, it must be that $\underline{b}^y_{M,s} \leq \underline{b}^y_{I,s}$ (and also that $\bar{b}^y_{M,s} \leq \bar{b}^y_{I,s}$). First, we note that once $x \geq \theta_s - \underline{b}^y_{M,s}$ the agent gets admitted with probability 1 with either M or I, and therefore would choose the cheaper action. This means that for $\theta_s - \underline{b}^y_{M,s} \leq x < \theta_s$, the agent chooses action M over I. Given the continuity of the utility functions under actions M and I, we expect this argument to carry for some x smaller than $\theta_s - \underline{b}^y_{M,s}$ as well.

Specifically, once $\theta_s - \underline{b}^y_{I,s} \leq x \leq \theta_s - \underline{b}^y_{M,s}$, the agent receives sufficient boost to get admitted with probability 1 when choosing action I, but is still uncertain when choosing M. Formally, improvement has a utility of $1 - C_{I,s}$, whereas manipulation has a utility of $1 - T^y_{M,s}(\theta_s - x) - C_{M,s}$. Therefore, if $1 - T^y_{M,s}(\theta_s - x) - C_{M,s} \geq 1 - C_{I,s}$ in this region, or equivalently once $x \geq \mathbf{r}^y_s$, the uncertainty from action M is small enough for the agent to choose action M over I. Note also that this argument will continue to hold even if the indifference point $\mathbf{r}^y_s$ is such that $\mathbf{r}^y_s \leq \theta_s - \underline{b}^y_{I,s}$, because this would only increase the uncertainty about I, making the utility from choosing I smaller than $1 - C_{I,s}$; this means that the utility of action M will still be higher than the utility of I when $x \geq \mathbf{r}^y_s$, making action M preferable to I.

Finally, agents with $\max\{\mathbf{o}^y_{I,s}, \mathbf{o}^y_{M,s}\} \leq x \leq \min\{\mathbf{r}^y_s, \theta_s - \underline{b}^y_{I,s}\}$ (provided the region is non-empty) would benefit from either manipulation or improvement actions over doing nothing, but face uncertainties about making it to an admit decision when opting for these actions, leading to a cost-efficacy trade off between these choices. Formally, define $\Delta u_s(x, y) := u_s(x, y, I) - u_s(x, y, M)$, the difference between the utility of improvement and manipulation. This difference is:

$$\begin{aligned} \Delta u_s(x, y) &= \mathbb{P}(\hat{x} \geq \theta_s | X = x, Y = y, W = M, S = s) - C_{M,s} \\ &\quad - (\mathbb{P}(\hat{x} \geq \theta_s | X = x, Y = y, W = I, S = s) - C_{I,s}) \\ &= \left(\mathbb{T}^y_{M,s}(\theta_s - x) - \mathbb{T}^y_{I,s}(\theta_s - x)\right) - (C_{I,s} - C_{M,s}) \ . \end{aligned} \tag{9}$$

If this difference is positive, the agent will opt for I over M. Recall that $\mathbf{f}^y_s$ is the feature such that $\mathbb{T}^y_{M,s}(\theta_s - \mathbf{f}^y_s) - \mathbb{T}^y_{I,s}(\theta_s - \mathbf{f}^y_s) = C_{I,s} - C_{M,s}$. This is the point at which the agent is indifferent between the M and I actions. If M is preferred to I before this point, this is the point at which the agent would switch from M to I, once I has a non-negative utility (and vice versa for when I is initially preferred to M).

Using the above characterizations, together with Lemma 6, leads to the identified best responses in each case.

B. Proof Lemma 1

We begin by noting that the new set of qualified ($\hat{Y} = 1$) agents consists of previously qualified agents, as well as previously unqualified agents who opted for improvement decisions. Therefore, the new qualification rate is given by $\hat{\alpha}_s = \alpha_s + (1 - \alpha_s) \int_{\mathbb{I}^0_s} G^0_s(x) \mathrm{d}x$.

Next, note that the set of (now) qualified agents with feature x consists of the previously qualified agents with the same old feature x (who have opted for $w = N$), the previously qualified or unqualified agents with feature $x - b$ who improved to feature x (i.e., opted for $w = I$ and got a boost realization b), and previously qualified agents with feature $x - b$ who chose manipulation and reached feature x (i.e., opted for $w = M$ and got a boost realization b). Thus,

$$\begin{aligned} \hat{G}^1_s(x) = \tfrac{\alpha_s}{\hat{\alpha}_s} \big(&\mathbb{1}(x \in \mathbb{N}^1_s) G^1_s(x) + \int_b \mathbb{1}(x - b \in \mathbb{M}^1_s) G^1_s(x - b) \\ &\tau^1_{M,s}(b) \mathrm{d}b \quad + \int_b \mathbb{1}(x - b \in \mathbb{I}^1_s) G^1_s(x - b) \tau^1_{I,s}(b) \mathrm{d}b \big) \\ &+ \tfrac{1 - \alpha_s}{\hat{\alpha}_s} \int_b \mathbb{1}(x - b \in \mathbb{I}^0_s) G^0_s(x - b) \tau^0_{I,s}(b) \mathrm{d}b \end{aligned} \tag{10}$$

We can re-write the expression for the integrals over the boost values using a change of variable $z := x - b$ as follows:

$$\begin{aligned} &\int_0^\infty \mathbb{1}(x - b \in \mathbb{M}^1_s) G^1_s(x - b) \tau^1_{M,s}(b) \mathrm{d}b \\ &= \int_x^{-\infty} \mathbb{1}(z \in \mathbb{M}^1_s) G^1_s(z) \tau^1_{M,s}(x - z) \mathrm{d}(x - z) \\ &= \int_{-\infty}^{x} \mathbb{1}(z \in \mathbb{M}^1_s) G^1_s(z) \tau^1_{M,s}(x - z) \mathrm{d}z \\ &= \int_{z \in \mathbb{M}^1_s} G^1_s(z) \tau^1_{M,s}(x - z) \mathrm{d}z \ . \end{aligned}$$

Substituting the above in (10) leads to (6).

Using similar arguments, the post-strategic feature distribution of (now) unqualified agents from group s consists of unqualified agents with the same feature who have opted to do nothing, and previously unqualified agents who have opted for manipulation and have reached feature x. Thus:

$$\begin{aligned} \hat{G}^0_s(x) = \tfrac{1 - \alpha_s}{1 - \hat{\alpha}_s} \big(&\mathbb{1}(x \in \mathbb{N}^0_s) G^0_s(x) \\ &+ \int_b \mathbb{1}(x - b \in \mathbb{M}^0_s) G^0_s(x - b) \tau^0_{M,s}(b) \mathrm{d}b \big) \end{aligned} \tag{11}$$

Re-writing the integral similar to before leads to (5).