The Double-Edged Sword of Behavioral Responses in Strategic Classification

Raman Ebrahimi
ECE, UC San Diego
raman@ucsd.edu

Kristen Vaccaro
CSE, UC San Diego
kv@ucsd.edu

Parinaz Naghizadeh
ECE, UC San Diego
parinaz@ucsd.edu

Abstract

When humans are subject to an algorithmic decision system, they can choose to strategically adjust their behavior accordingly ("game" the system). While a growing line of literature on strategic classification has used game-theoretic modeling to understand and mitigate such gaming, these existing works consider standard models of *fully rational* agents. In this paper, we propose a model of strategic classification which takes into account *behavioral biases* in human responses to algorithms. We show how misperceptions of the classifier (specifically, of its feature weights) can lead to different types of discrepancies between biased and rational agents' gaming responses, and identify when behavioral agents over- or under-investment in different features. We also show that strategic agents with behavioral biases can benefit or (perhaps, unexpectedly) harm the firm compared to fully rational strategic agents, highlighting the need to account for human (cognitive) biases when designing AI systems with strategic human in the loop.

1 Introduction

As machine learning systems become more widely deployed, particularly in settings such as resume screening, hiring, lending, and recommendation systems, people have begun to respond to them strategically. Often, this takes the form of "gaming the system" or using an algorithmic system's own rules and procedures in order to manipulate it and achieve desired outcomes. Examples include Uber drivers coordinating the times they log on and off the app to impact its surge pricing algorithm (Möhlmann and Zalmanson, 2017), and Twitter (Burrell et al., 2019) and Facebook (Eslami et al., 2016) users' decisions regarding how to interact with content given the platforms' curation algorithms.

Game theoretical modeling and analysis have been used in recent years to formally analyze such strategic responses of humans to algorithms (e.g., Hardt et al. (2016); Milli et al. (2019); Liu et al. (2020); see also Related Work). However, these existing works assume *standard* models of decision making, where agents are fully rational when responding to algorithms; yet, humans exhibit different forms of cognitive biases in decision making (Kahnemann and Tversky, 1979). Motivated by this, in this paper, we explore the impacts *behavioral biases* on agents' strategic responses to algorithms.

We begin by proposing an extension of existing models of strategic classification to account for behavioral biases. Specifically, our model accounts for agents misperceiving (e.g., over-weighing or under-weighing) the importance of different features in determining the classifier's output. (These may be known to agents in a full information game or can become available through an explainable AI (XAI) method). We use this model to identify different forms of discrepancies that can arise between behavioral and fully rational agents' responses (Section 3). We further identify conditions under which agents' behavioral biases lead them to over- or under-invest in specific features (Proposition 2). Moreover, we show that a firm's utility could increase or decrease when agents are behaviorally biased, compared to when they are fully rational (Proposition 3). While the former may be intuitively expected (behaviorally biased agents are less adept at gaming algorithms), the latter is more surprising;

NeurIPS 2024 Workshop on Behavioral Machine Learning.

we intuitively explain this through a numerical example (Example 1) and by highlighting the impact of agents' qualification states in determining the ultimate impact of agents' behavioral biases on the firm. Our findings highlight the necessity to account for not just strategic responses, but also cognitive biases, when designing AI systems with human in the loop.

Related Work. Our work is closely related to the literature on analyzing agents' responses to machine learning algorithms, when agents have full (Hardt et al., 2016; Perdomo et al., 2020; Milli et al., 2019; Hu et al., 2019; Liu et al., 2020; Bechavod et al., 2022; Kleinberg and Raghavan, 2020; Alhanouti and Naghizadeh, 2024; Zhang et al., 2022; Bechavod et al., 2021) or partial information about the algorithm (Harris et al., 2022; Cohen et al., 2024), or principal's strategy (Haghtalab et al., 2024). While our base model of agents' strategic responses to classifiers has similarities to those in some of these works (e.g., Hu et al. (2019); Liu et al. (2020)), we differ in our modeling of agent's *behavioral* responses as opposed to fully *rational* (non-behavioral) best responses considered in these works.

The necessity of accounting for human biases in making AI assisted decisions (Rastogi et al., 2022; Nourani et al., 2021), and various aspects of decision-making and model design (Morewedge et al., 2023; Zhu et al., 2024; Liu et al., 2024; Heidari et al., 2021; Ethayarajh et al., 2024) have been considered in recent work. Among these, Heidari et al. (2021) uses probability weighting functions to model human perceptions of allocation policies. We also consider (Prelec) weighting functions, but to highlight special cases of our results. We also differ from all these existing works in our focus on the *strategic classification* problem.

Broadly, our research is also related to the area of explainable AI. While explanations can be helpful in increasing accountability, there is debate about the efficacy of existing explainability methods in providing correct and sufficient details in a way that helps users understand and act around these systems (Doshi-Velez et al., 2017; Kumar et al., 2020; Lakkaraju and Bastani, 2020; Adebayo et al., 2018). To complement these discussions, our work provides a formal model of how agents' behavioral biases may shape their responses to explanations (of feature importance) provided to them.

2 Model and Preliminaries

Strategic Classification. We consider an environment in which a *firm* makes binary classification decisions on *agents* with (observable) features $\mathbf{x} \in \mathbb{R}^n$ and (unobservable) true qualification states/labels $y \in \{0, 1\}$, where label $y = 1$ (resp. $y = 0$) denotes qualified (resp. unqualified) agents. The firm uses a threshold classifier $h(\boldsymbol{x}, (\boldsymbol{\theta}, \theta_0)) = \mathbf{1}(\boldsymbol{\theta}^T \boldsymbol{x} \geq \theta_0)$ to classify agents, where $\mathbf{1}(\cdot)$ denotes the indicator function, and $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_n]^T$ denotes the *feature weights*; *w.l.o.g.*, we assume features are indexed such that $\theta_1 \geq \theta_2 \geq \dots \theta_n$, and are normalized so that $\sum_i \theta_i = 1$. Agents are strategic, in that they can respond to ("game") this classifier. (As an example, in a college admission setting where grades are used to make admission decisions, students can study or cheat to improve their grades.) Formally, an agent with *pre-strategic* features $\boldsymbol{x}_0$ best-responds to classifier $(\boldsymbol{\theta}, \theta_0)$ to arrive at the *(non-behavioral) post-strategic* features $\boldsymbol{x}_{\text{NB}}$ by solving the optimization problem $\boldsymbol{x}_{\text{NB}} := \arg\max_{\boldsymbol{x}} \ rh(\boldsymbol{x}, (\boldsymbol{\theta}, \theta_0)) - c(\boldsymbol{x}, \boldsymbol{x}_0)$ subject to $c(\boldsymbol{x}, \boldsymbol{x}_0) \leq B$, where $r > 0$ is the reward of positive classification, $c(\boldsymbol{x}, \boldsymbol{x}_0)$ is *norm-2 cost* (with $c(\boldsymbol{x}, \boldsymbol{x}_0) = \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 = \sum_i (x_i - x_{i,0})^2$) of changing feature vector $\boldsymbol{x}_0$ to $\boldsymbol{x}$, and B is the agent's budget.

Anticipating agents' best-response, the firm will choose the optimal (non-behavioral) classifier threshold by solving $(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}) := \arg\min_{(\boldsymbol{\theta}, \theta_0)} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}(\boldsymbol{\theta}, \theta_0)}[l(\boldsymbol{x}, (\boldsymbol{\theta}, \theta_0))]$, where $\mathcal{D}(\boldsymbol{\theta}, \theta_0)$ is the post-strategic feature distribution of agents responding to classifier $(\boldsymbol{\theta}, \theta_0)$ and $l(\cdot)$ is the firm's loss function (e.g., a weighted sum of true positive and false positive costs).

Behavioral Responses. We extend the above problem setting to allow for behavioral responses by agents. Formally, recall that we normalize the feature weight vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_n]^T$ to have $\sum_i \theta_i = 1$, and interpret it as a probability vector. Given this, we will assume that behaviorally-biased agents misperceive $\boldsymbol{\theta}$ as $\boldsymbol{w}(\boldsymbol{\theta})$, where $\boldsymbol{w}(\cdot)$ is a function capturing their behavioral biases. One choice for $\boldsymbol{w}(\cdot)$ can be $w_j(\boldsymbol{\theta}) = p(\sum_{i=1}^{j} \theta_i) - p(\sum_{i=1}^{j-1} \theta_i)$, to ensure $\sum_i w_i(\boldsymbol{\theta}) = 1$ (Gonzalez and Wu, 1999) where $p(z) = \exp(-(-\ln(z))^\gamma)$ is the widely used probability weighting function introduced by Prelec (1998) with γ reflecting the intensity of biases.

Now, a behaviorally biased agent with pre-strategic features $\boldsymbol{x}_0$ best-responds to classifier $(\boldsymbol{\theta}, \theta_0)$ to arrive at the *behavioral post-strategic* features $\boldsymbol{x}_{\text{B}}$ by solving the optimization problem $\boldsymbol{x}_{\text{B}} := \arg\max_{\boldsymbol{x}} \ rh(\boldsymbol{x}, (\boldsymbol{w}(\boldsymbol{\theta}), \theta_0)) - c(\boldsymbol{x}, \boldsymbol{x}_0)$ subject to $c(\boldsymbol{x}, \boldsymbol{x}_0) \leq B$; note that the agent now responds to a *perceived decision boundary* $(\boldsymbol{w}(\boldsymbol{\theta}), \theta_0)$. In return, when accounting for agents' strategic behavior, the firm may or may not be aware that agents have behavioral biases. Specifically, let $\mathbb{L}(\boldsymbol{\theta}', (\boldsymbol{\theta}, \theta_0)) := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}(\boldsymbol{\theta}',\theta_0)}[l(\boldsymbol{x}, (\boldsymbol{\theta}, \theta_0))]$ denote a firm's expected loss when it implements a classifier $(\boldsymbol{\theta}, \theta_0)$ and agents respond to a classifier $(\boldsymbol{\theta}', \theta_0)$ (note that this is potentially different from the actual classifier). Then, if a firm is aware of strategic agents' behavioral biases, it selects the threshold $(\boldsymbol{\theta}_{\mathrm{B}}, \theta_{0,\mathrm{B}}) := \arg\min_{(\boldsymbol{\theta},\theta_0)} \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}), (\boldsymbol{\theta}, \theta_0))$ and incurs a loss $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_B), (\boldsymbol{\theta}_B, \theta_{0,\mathrm{B}}))$. On the other hand, a firm that assumes agents are fully rational selects the threshold classifier $(\boldsymbol{\theta}_{\mathrm{NB}}, \theta_{0,\mathrm{NB}})$, yet incurs the loss $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\mathrm{NB}}), (\boldsymbol{\theta}_{\mathrm{NB}}, \theta_{0,\mathrm{NB}}))$.

3 Fully Rational vs. Behavioral Best-Responses by Agents

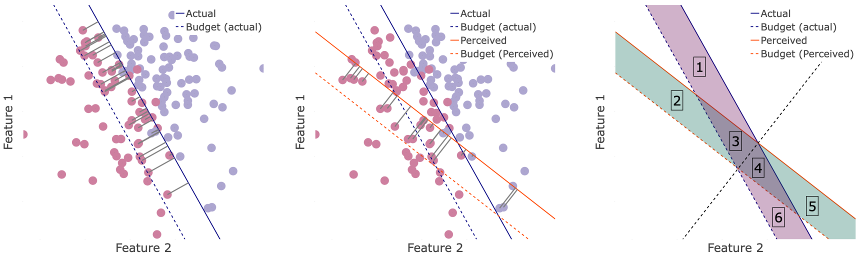
We first fix the classifier $(\boldsymbol{\theta}, \theta_0)$, and compare fully rational (non-behavioral) and behavioral agents' strategic responses to it. The following Lemma characterizes $\boldsymbol{x}_{\mathrm{NB}}$ and $\boldsymbol{x}_{\mathrm{B}}$ under the norm-2 cost.

Lemma 1. *Let $d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0) = \frac{\theta_0 - \boldsymbol{\theta}^T\boldsymbol{x}_0}{\|\boldsymbol{\theta}\|_2}$ denote $\boldsymbol{x}_0$'s distance to the hyperplane $\boldsymbol{\theta}^T\boldsymbol{x} = \theta_0$. Then, for an agent with starting feature vector $\boldsymbol{x}_0$, if $0 < d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0) \leq B$,*

$$\boldsymbol{x}_{NB} = \boldsymbol{x}_0 + d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)\boldsymbol{\theta} \ .$$

Otherwise, $\boldsymbol{x}_{NB} = \boldsymbol{x}_0$. For behaviorally biased agents, $\boldsymbol{x}_B$ is obtained similarly by replacing $\boldsymbol{\theta}$ with $\boldsymbol{w}(\boldsymbol{\theta})$.

Figure 1 illustrates the strategic agents' best-responses of Lemma 1, in a two-dimensional feature space, when they are non-behavioral (Fig. 1a) and when they are behavioral (Fig. 1b). We first note that the subset of agents with non-trivial responses to the classifier, as identified in Lemma 1, are in a band below the decision boundary. Given the overlaps of these bands under non-behavioral and behavioral responses, there are 6 regions of interest where biased agents' best-responses defer from rational agents (Fig. 1c). In regions $\boxed{1}$ and $\boxed{6}$, agents invest no effort in manipulating their features when they are behaviorally biased, whereas they do when fully rational; the reasons differ: agents in $\boxed{1}$ believe they are accepted without effort, while those in $\boxed{6}$ believe they do not have sufficient budget to succeed. Agents in regions $\boxed{2}$ and $\boxed{5}$ manipulate their features unnecessarily (they would not, had they been fully rational), and again, for different reasons: agents in $\boxed{2}$ are not accepted even at their highest effort level, while those in $\boxed{5}$ believe they must reach the boundary but they would be accepted regardless of their effort. Finally, in region $\boxed{3}$, agents *undershoot* the actual boundary (i.e., exert less effort than needed due to their biases), while those in region $\boxed{4}$, they *overshoot* (i.e., exert more effort than needed to get accepted). In the remainder of this section, we characterize agents that fall within certain regions (Proposition 1) and identify which features they over- or under-invest in (Proposition 2).



(a) Rational response (b) Biased response (c) Differing responses

Figure 1: (a) Fully rational and (b) Biased responses. (c) Classes of differing actions. Blue data points represent qualified agents and red data points represent unqualified agents.

We begin by characterizing the set of agents who fall within regions $\boxed{1}$ and $\boxed{3}$. These are the set of agents who will still pass the (true) decision boundary regardless of their biases.

Proposition 1. *For a given $(\boldsymbol{\theta}, \theta_0)$, agents that satisfy $(1 - \sigma(\boldsymbol{\theta}))\theta_0 \leq (\boldsymbol{\theta} - \sigma(\boldsymbol{\theta})\boldsymbol{w}(\boldsymbol{\theta}))^T\boldsymbol{x}$ will be accepted by the classifier, where $\sigma(\boldsymbol{\theta}) := \frac{\boldsymbol{\theta}^T\boldsymbol{w}(\boldsymbol{\theta})}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2}$ is a measure of the intensity of behavioral bias.*

Next, in the following proposition, we further investigate best-responses in region $\boxed{4}$ (resp. region $\boxed{3}$) and identify which features a behavioral agent over-invest (resp. under-invest) in that leads to them overshooting (resp. undershooting) past the true classifier $(\boldsymbol{\theta}, \theta_0)$.

Proposition 2. *Consider an agent with features $\boldsymbol{x}_0$, facing classifier $(\boldsymbol{\theta}, \theta_0)$, and with a misperceived $\boldsymbol{w}(\boldsymbol{\theta})$. Let $\theta_{\max} = \max_i \theta_i$, $d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0) = \frac{\theta_0 - \boldsymbol{\theta}^T \boldsymbol{x}_0}{\|\boldsymbol{\theta}\|_2}$, and $\delta_i^{NB} = x_{NB,i} - x_{0,i}$ and $\delta_i^{B} = x_{B,i} - x_{0,i}$ denote the changes in feature i after best-responses. Then:*

***(1)** If $d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0) \leq d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)$ and $w(\theta_i) < \theta_i$, then $\delta_i^{B} < \delta_i^{NB}$.*

***(2)** If $d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0) \leq d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0)$ and $\theta_i < w(\theta_i)$ then $\delta_i^{NB} < \delta_i^{B}$.*

***(3)** For the special case of a Prelec function, we further have: If $d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0) \leq e^{\gamma^{\frac{1}{1-\gamma}} - \gamma^{\frac{\gamma}{1-\gamma}}} d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0)$ and $w(\theta_{\max}) < \theta_{\max}$, then $\delta_{\max}^{NB} < \delta_{\max}^{B}$.*

Intuitively, the proposition states that agents that perceive the decision boundary to be closer to them than it truly is (regions $\boxed{2}$ and $\boxed{3}$ in Figure 1c) will under-invest in the features for which they underestimate the importance. Similarly, agents that perceive the boundary to be farther (regions $\boxed{4}$ and $\boxed{5}$ in Figure 1c) will over-invest in the features for which they overestimate the importance.

4 Firm's Response

We next consider the firm's optimal choice of a classifier, given agents' strategic responses, and its impact on the firm's utility. Intuitively, one might expect a firm to ultimately benefit from agents' behavioral responses (in contrast to fully rational responses) as behavioral agents are less adept at gaming the algorithm. However, in this section, we show that this is not always true. Intuitively, as shown in Section 3, when behavioral, agents may overshoot or undershoot the threshold when gaming the algorithm; this includes both qualified (label 1) and unqualified (label 0) agents. We show that there exist scenarios in which a relatively higher number of behaviorally biased qualified agents end up below the threshold (due to not trying or undershooting) while relatively more unqualified agents overshoot and end up accepted by the classifier; the combination of these factors can lower the firm's utility. The following example numerically illustrates this possibility.

Example 1. *Consider a setting where we have a two-dimensional feature space and qualified (resp. unqualified) agents are sampled from a normal distribution $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ (resp. $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$). We consider two scenarios, only differing in the mean $\boldsymbol{\mu}_1$ choice. Figure 2 illustrates the distribution of agents' features for pre-strategic (left panel), post-strategic non-behavioral responses (middle panel), and post-strategic behaviorally-biased responses (right panel). In the top row scenario, the firm is negatively impacted by agents' bias, while in the bottom row scenario, the firm benefits from agents' bias compared to the fully rational setting. (The firm's payoff in each case is shown at the top of the corresponding subplots.) The reason for this difference is that there are more qualified agents than unqualified ones who reach the threshold in non-biased responses. On the other hand, under biased responses, there are more unqualified agents who pass the threshold, regardless of their bias (those in region $\boxed{3}$ in Fig. 1c) in the top row scenario. Behavioral responses by these agents negatively impact the firm, as it leads to these qualified agents no longer being accepted.*

The following proposition formalizes the intuition from Example 1.

Proposition 3. *Consider a loss function $l(\boldsymbol{x}, (\boldsymbol{\theta}, \theta_0)) = -u^{+}TP + u^{-}FP$. Let the pdf of label y agents' feature distribution be $f_y(\boldsymbol{x})$, and the number of label y agents be α_0. Let $\mathcal{H}(\boldsymbol{\theta}, \theta_0)$ denote the set of agents that satisfy $(1 - \sigma(\boldsymbol{\theta}))\theta_0 \leq (\boldsymbol{\theta} - \sigma(\boldsymbol{\theta})\boldsymbol{w}(\boldsymbol{\theta}))^T \boldsymbol{x}$, where $\sigma(\boldsymbol{\theta}) := \frac{\boldsymbol{\theta}^T \boldsymbol{w}(\boldsymbol{\theta})}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2}$[1], and the set of agents that attempt to game the algorithm as $\mathbb{A}(\boldsymbol{\theta}, \theta_0) = \{\boldsymbol{x}_0 : \theta_0 - B \leq \boldsymbol{\theta}^T \boldsymbol{x}_0 < \theta_0\}$. Denote the set of accepted (resp. rejected) agents by $(\boldsymbol{\theta}, \theta_0)$ with $\mathbb{Y}(\boldsymbol{\theta}, \theta_0)$ (resp. $\mathbb{N}(\boldsymbol{\theta}, \theta_0)$). Define the sets $\mathbb{S}(\boldsymbol{\theta}_{NB}, \theta_{0,NB}) := \mathbb{A}(\boldsymbol{\theta}_{NB}, \theta_{0,NB}) / (\mathbb{A}(\boldsymbol{\theta}_{NB}, \theta_{0,NB}) \cap \mathcal{H}(\boldsymbol{\theta}_{NB}, \theta_{0,NB}))$, $\mathbb{T}_1 = (\mathbb{Y}(\boldsymbol{\theta}_{NB}, \theta_{0,NB}) \cup \mathbb{A}(\boldsymbol{\theta}_{NB}, \theta_{0,NB})) \cap \mathbb{N}(\boldsymbol{\theta}_B, \theta_{0,B})$, and $\mathbb{T}_2 = (\mathcal{H}(\boldsymbol{\theta}_B, \theta_{0,B}) \cap \mathbb{A}(\boldsymbol{w}(\boldsymbol{\theta}_B), \theta_{0,B})) \cup ((\mathbb{Y}(\boldsymbol{\theta}_B, \theta_{0,B}) \cap \mathbb{N}(\boldsymbol{\theta}_{NB}, \theta_{0,NB})) / \mathbb{A}(\boldsymbol{\theta}_{NB}, \theta_{0,NB}))$. Then:*

***(a)** If $\int_{x \in \mathbb{S}(\boldsymbol{\theta}_{NB}, \theta_{0,NB})} u^{-} f_0(\boldsymbol{x}) \alpha_0 d\boldsymbol{x} \leq \int_{x \in \mathbb{S}(\boldsymbol{\theta}_{NB}, \theta_{0,NB})} u^{+} f_1(\boldsymbol{x}) \alpha_1 d\boldsymbol{x}$ we can say:*

$$\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_B), (\boldsymbol{\theta}_B, \theta_{0,B})) \leq \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{NB}), (\boldsymbol{\theta}_{NB}, \theta_{0,NB})) \leq \mathbb{L}(\boldsymbol{\theta}_{NB}, (\boldsymbol{\theta}_{NB}, \theta_{0,NB})) \tag{1}$$

[1] Note that $\sigma(\boldsymbol{\theta}) = \frac{\|\boldsymbol{\theta}\|_2}{\|\boldsymbol{w}(\boldsymbol{\theta})\|_2} \cos(\alpha)$ where α is the angle between the actual and perceived decision boundaries. The larger α is, the lower $\sigma(\boldsymbol{\theta})$ is, indicating a more intense bias.

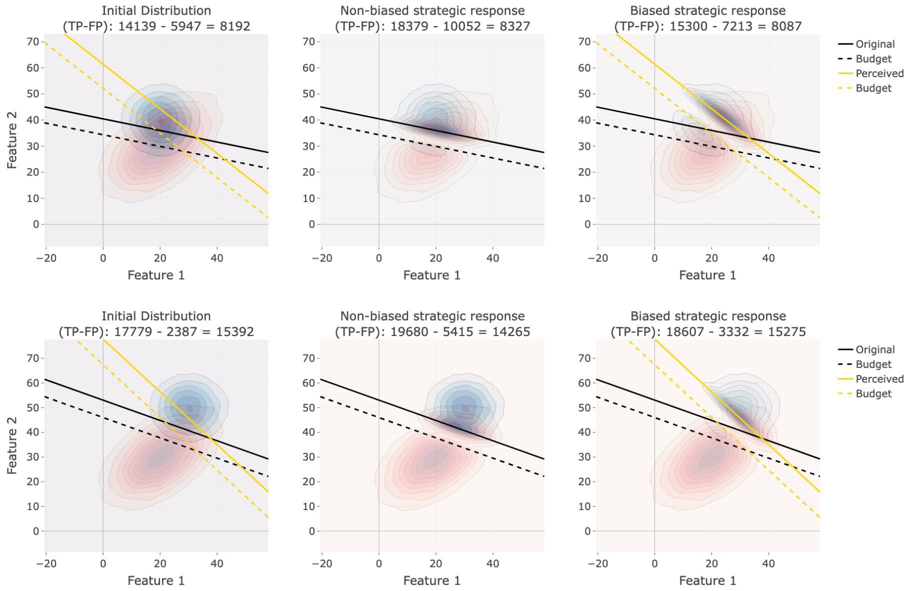


Figure 2: The firm may have lower (top) or higher (bottom) utility when agents are behavioral. Blue represents the distribution of qualified agents and red represents the distribution of unqualified agents.

(b) *If* $\int_{x\in\mathbb{S}(\boldsymbol{\theta}_{NB},\theta_{0,NB})} u^+ f_1(\boldsymbol{x})\alpha_1 d\boldsymbol{x} \le \int_{x\in\mathbb{S}(\boldsymbol{\theta}_{NB},\theta_{0,NB})} u^- f_0(\boldsymbol{x})\alpha_0 d\boldsymbol{x}$ *we can say:*

$$\max\{\mathbb{L}(\boldsymbol{\theta}_{NB},(\boldsymbol{\theta}_{NB},\theta_{0,NB})),\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_B),(\boldsymbol{\theta}_B,\theta_{0,B}))\} \le \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{NB}),(\boldsymbol{\theta}_{NB},\theta_{0,NB})) \tag{2}$$

(c) *If* $\int_{\boldsymbol{x}\in\mathbb{T}_1}(-u^+ f_1(\boldsymbol{x})\alpha_1 + u^- f_0(\boldsymbol{x})\alpha_0)d\boldsymbol{x} \le \int_{\boldsymbol{x}\in\mathbb{T}_2}(-u^+ f_1(\boldsymbol{x})\alpha_1 + u^- f_0(\boldsymbol{x})\alpha_0)d\boldsymbol{x}$ *we can say:*

$$\mathbb{L}(\boldsymbol{\theta}_{NB},(\boldsymbol{\theta}_{NB},\theta_{0,NB})) \le \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_B),(\boldsymbol{\theta}_B,\theta_{0,B})) \le \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{NB}),(\boldsymbol{\theta}_{NB},\theta_{0,NB})) \tag{3}$$

Part (a) states that if a firm is unaware of agents' behavioral biases, it will suffer a lower loss when the population is biased compared to fully rational. This is the intuitively expected scenario (behaviorally biased agents are less adept than fully rational ones at gaming the algorithm). On the other hand, statement (b) reflects the less expected outcome: a firm unaware of behavioral biases will have *lower* utility when agents are biased compared to if they had been fully rational (as more *qualified* than *unqualified* agents undershoot the threshold under this case's condition). Statement (c) further compares the unaware firm with an aware firm and provides a condition where an aware firm's minimal loss is higher than the non-biased minimal loss. This condition relies on the *difference* of qualified and unqualified agents in two regions.

5 Conclusion

We present a strategic classification framework that accounts for the cognitive biases of strategic agents when assessing feature importance. We identify conditions under which the agents over- or under-invest in different features, the impacts of this on a firm's choice of classifier, and the impacts on the firm's utility. Exploring analytical models accounting for biases beyond misperception of feature weights, investigating fairness implications when these biases differ across demographic groups, as well as providing support for, and measuring, the existence of these biases through human subject experiments, remain as important directions for further investigation.

Acknowledgments and Disclosure of Funding

This work was supported in part by the UCSD Jacobs School Early Career Faculty Development Award, and by the NSF program on Fairness in AI in collaboration with Amazon under Award IIS-2040800. Any opinions and recommendations expressed in this material are those of the authors.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9525–9536, Red Hook, NY, USA. Curran Associates Inc.

Alhanouti, S. and Naghizadeh, P. (2024). Could anticipating gaming incentivize improvement in (fair) startegic classification? *The IEEE Control and Decisions Conference (CDC)*.

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*.

Bechavod, Y., Podimata, C., Wu, S., and Ziani, J. (2022). Information discrepancy in strategic learning. In *International Conference on Machine Learning*, pages 1691–1715. PMLR.

Burrell, J., Kahn, Z., Jonas, A., and Griffin, D. (2019). When users control the algorithms: values expressed in practices on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20.

Cohen, L., Sharifi-Malvajerdi, S., Stangl, K., Vakilian, A., and Ziani, J. (2024). Bayesian strategic classification. *arXiv preprint arXiv:2402.08758*.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., et al. (2017). Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., and Kirlik, A. (2016). 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*, CHI '16, page 2371–2382.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. (2024). Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Gonzalez, R. and Wu, G. (1999). On the shape of the probability weighting function. *Cognitive psychology*, 38(1):129–166.

Haghtalab, N., Podimata, C., and Yang, K. (2024). Calibrated stackelberg games: Learning optimal commitments against calibrated agents. *Advances in Neural Information Processing Systems*, 36.

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*, ITCS '16, page 111–122, New York, NY, USA. Association for Computing Machinery.

Harris, K., Chen, V., Kim, J., Talwalkar, A., Heidari, H., and Wu, S. Z. (2022). Bayesian persuasion for algorithmic recourse. *Advances in Neural Information Processing Systems*, 35:11131–11144.

Heidari, H., Barocas, S., Kleinberg, J., and Levy, K. (2021). On modeling human perceptions of allocation policies with uncertain outcomes. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, page 589–609, New York, NY, USA. Association for Computing Machinery.

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*, FAT* '19, page 259–268, New York, NY, USA. Association for Computing Machinery.

Kahnemann, D. and Tversky, A. (1979). Prospect theory: A decision making under risk. *Econometrica*, 47(2):263–291.

Kleinberg, J. and Raghavan, M. (2020). How do classifiers induce agents to invest effort strategically? *ACM Trans. Econ. Comput.*, 8(4).

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. (2020). Problems with shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Lakkaraju, H. and Bastani, O. (2020). "how do i fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 79–85, New York, NY, USA. Association for Computing Machinery.

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*, FAT* '20, page 381–391, New York, NY, USA. Association for Computing Machinery.

Liu, R., Geng, J., Peterson, J. C., Sucholutsky, I., and Griffiths, T. L. (2024). Large language models assume people are more rational than we really are. *arXiv preprint arXiv:2406.17055*.

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*, FAT* '19, page 230–239, New York, NY, USA. Association for Computing Machinery.

Möhlmann, M. and Zalmanson, L. (2017). Hands on the wheel: Navigating algorithmic management and uber drivers'. In *Autonomy', in proceedings of the international conference on information systems (ICIS), Seoul South Korea*, pages 10–13.

Morewedge, C. K., Mullainathan, S., Naushan, H. F., Sunstein, C. R., Kleinberg, J., Raghavan, M., and Ludwig, J. O. (2023). Human bias in algorithm design. *Nature Human Behaviour*, 7(11):1822–1824.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., and Gogate, V. (2021). Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 340–350.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3):497–527.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., and Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22.

Zhang, X., Khalili, M. M., Jin, K., Naghizadeh, P., and Liu, M. (2022). Fairness interventions as (Dis)Incentives for strategic manipulation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26239–26264. PMLR.

Zhu, J.-Q., Peterson, J. C., Enke, B., and Griffiths, T. L. (2024). Capturing the complexity of human strategic decision-making with machine learning. *arXiv preprint arXiv:2408.07865*.

A Proofs

Proof of Lemma 1 We show the NB case, the B case can be shown similarly. We divide the agents into two subsets: (1) Agents that will attempt to optimize and (2) agents that will not attempt to optimize. The first subset is the agents that will have a non-negative utility after optimization, i.e., will have $r - c(\boldsymbol{x}_{\text{NB}}, \boldsymbol{x}_0)$. For these agents, since their reward is constant, the optimization problem comes down to:

$$\begin{aligned} \boldsymbol{x}_{\text{NB}} &:= \arg\max_{\boldsymbol{x}} \; r - c(\boldsymbol{x}, \boldsymbol{x}_0) \\ &\text{subject to} \quad \boldsymbol{\theta}^T \boldsymbol{x} = \theta_0 \end{aligned} \tag{4}$$

And the agents that are in the second subset will solve $\boldsymbol{x}_{\text{NB}} := \arg\min_{\boldsymbol{x}} \; c(\boldsymbol{x}, \boldsymbol{x}_0)$ which is simply $\boldsymbol{x}_{\text{NB}} = \boldsymbol{x}_0$.

For norm-2 cost we know this is the same as finding the closest point on a hyperplane to a given point. We know the solution for this problem is to move in the direction of the normal vector of the hyperplane by $d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0) = \frac{\theta_0 - \boldsymbol{\theta}^T \boldsymbol{x}_0}{\|\boldsymbol{\theta}\|_2}$. This means that the solution for the agents in the first subset is $\boldsymbol{x}_{\text{NB}} = \boldsymbol{x}_0 + d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)\boldsymbol{\theta}$.

Proof of Proposition 1 We can write agents' behavioral response as $\boldsymbol{x} + \Delta_{\text{B}}$ with $\Delta_{\text{B}} = \frac{\theta_0 - \boldsymbol{w}(\boldsymbol{\theta})^T \boldsymbol{x}}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2} \boldsymbol{w}(\boldsymbol{\theta})$ for a given $(\boldsymbol{\theta}, \theta_0)$. Agents that will have successful manipulation are the ones satisfying $\theta_0 \leq \boldsymbol{\theta}^T(\boldsymbol{x} + \Delta_{\text{B}})$ which, by substituting Δ_{B}, can be written as:

$$\begin{aligned} &\boldsymbol{\theta}_0 \leq \frac{\theta_0 - \boldsymbol{w}(\boldsymbol{\theta})^T \boldsymbol{x}}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2} \boldsymbol{\theta}^T \boldsymbol{w}(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \boldsymbol{x} = \frac{\boldsymbol{\theta}^T \boldsymbol{w}(\boldsymbol{\theta})}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2} \theta_0 + \left(\boldsymbol{\theta} - \frac{\boldsymbol{\theta}^T \boldsymbol{w}(\boldsymbol{\theta})}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2} \boldsymbol{w}(\boldsymbol{\theta}) \right)^T \boldsymbol{x} \\ &\Rightarrow (1 - \sigma(\boldsymbol{\theta}))\theta_0 \leq (\boldsymbol{\theta} - \sigma(\boldsymbol{\theta})\boldsymbol{w}(\boldsymbol{\theta}))^T \boldsymbol{x} \qquad (5) \end{aligned}$$

Where we defined $\sigma(\boldsymbol{\theta}) := \frac{\boldsymbol{\theta}^T \boldsymbol{w}(\boldsymbol{\theta})}{\|\boldsymbol{w}(\boldsymbol{\theta}\|^2}$.

Proof of Proposition 2 For a behavioral agent with $\boldsymbol{x}_0$ that perceives θ_i as $w_i(\boldsymbol{\theta})$ to under-invest we need to have $\delta_i^{\text{B}} = d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0) \times w_i(\boldsymbol{\theta}) < \delta_i^{\text{NB}} = d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0) \times \theta_i$, or $\frac{d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0)}{d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)} < \frac{\theta_i}{w_i(\boldsymbol{\theta})}$.

By knowing $w_i(\boldsymbol{\theta}) < \theta_i$ then the agents with $d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0) \leq d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)$ will satisfy the condition since $\frac{d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0)}{d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)} \leq 1 < \frac{\theta_i}{w_i(\boldsymbol{\theta})}$ and under-invest in feature i. We can show the second statement similarly.

The third statement of the proposition is a scenario where $w_1(\theta) < \theta_1$ where $\theta_1 \geq \theta_i$ for all i, and we want to identify agents that will over-invest in that feature, i.e., $\frac{d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0)}{d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)} > \frac{\theta_1}{w_1(\boldsymbol{\theta})}$.

Since for the most important feature we have $w_1(\boldsymbol{\theta}) = p(\theta_1)$, we can easily find the maximum of $\frac{\theta_1}{w_1(\boldsymbol{\theta})}$ for a given γ by taking the derivative and using the function in Prelec (1998). This maximum occurs at $\theta^* = e^{-(\frac{1}{\gamma})^{\frac{1}{\gamma-1}}}$ meaning, $\frac{\theta_1}{w_1(\boldsymbol{\theta})} \leq \frac{\theta^*}{w(\theta^*)} = \exp\Big((\frac{1}{\gamma})^{\frac{\gamma}{\gamma-1}} - (\frac{1}{\gamma})^{\frac{1}{\gamma-1}} \Big)$. Therefore, using the same reasoning for the first two statements, agents with $\frac{d(\boldsymbol{x}_0, \boldsymbol{w}(\boldsymbol{\theta}), \theta_0)}{d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)} \geq \exp\Big((\frac{1}{\gamma})^{\frac{\gamma}{\gamma-1}} - (\frac{1}{\gamma})^{\frac{1}{\gamma-1}} \Big)$ will over-invest in the most important feature, i.e., feature 1.

Proof of Proposition 3 We start the proof from the leftmost inequality in equation 1. By the definition of $(\boldsymbol{\theta}_{\text{B}}, \boldsymbol{\theta}_{0,\text{B}})$ we can write $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{B}}), \theta_{0,\text{B}})}[l(\boldsymbol{x}, (\boldsymbol{\theta}_{\text{B}}, \boldsymbol{\theta}_{0,\text{B}}))] \leq \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}(\boldsymbol{w}(\boldsymbol{\theta}), \theta_0)}[l(\boldsymbol{x}, (\boldsymbol{\theta}, \boldsymbol{\theta}_0))]$ for all $(\boldsymbol{\theta}, \boldsymbol{\theta}_0) \neq (\boldsymbol{\theta}_{\text{B}}, \boldsymbol{\theta}_{0,\text{B}})$, i.e., $\mathbb{L}((\boldsymbol{w}(\boldsymbol{\theta}_{\text{B}}), \theta_{0,\text{B}}), (\boldsymbol{\theta}_{\text{B}}, \boldsymbol{\theta}_{0,\text{B}})) \leq \mathbb{L}((\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}), \theta_{0,\text{NB}}), (\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}))$ is always true.

To compare the firm's loss after biased and non-biased responses, we can break the feature space into the following regions ($\mathbf{1}(\cdot)$ is the indicator function):

$\textcircled{1}$ $\mathbf{1}(\boldsymbol{\theta}_{\text{NB}}^T \boldsymbol{x} \geq \theta_{0,\text{NB}})$

$\textcircled{2}$ $\mathbf{1}(\boldsymbol{\theta}_{\text{NB}}^T \boldsymbol{x} \leq \theta_{0,\text{NB}} - B)$

$\textcircled{3}$ $\mathbf{1}(\theta_{0,\text{NB}} - B \leq \boldsymbol{\theta}_{\text{NB}}^T \boldsymbol{x} \leq \theta_{0,\text{NB}}) \mathbf{1}(\theta_{0,\text{NB}} - B \leq \boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}})^T \boldsymbol{x} \leq \theta_{0,\text{NB}}) \equiv \mathbb{A}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}) \cap \mathbb{A}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}), \theta_{0,\text{NB}})$

$\textcircled{4}$ $\mathbf{1}(\theta_{0,\text{NB}} - B \leq \boldsymbol{\theta}_{\text{NB}}^T \boldsymbol{x} \leq \theta_{0,\text{NB}}) \mathbf{1}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}})^T \boldsymbol{x} \geq \theta_{0,\text{NB}})$

$\textcircled{5}$ $\mathbf{1}(\theta_{0,\text{NB}} - B \leq \boldsymbol{\theta}_{\text{NB}}^T \boldsymbol{x} \leq \theta_{0,\text{NB}}) \mathbf{1}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}})^T \boldsymbol{x} \leq \theta_{0,\text{NB}} - B)$

We know that for $\boldsymbol{x} \in \textcircled{1}$ and $\boldsymbol{x} \in \textcircled{2}$, the expected loss for both response scenarios is the same since the agents in the two regions are either already qualified or will never make it to the decision boundary. Therefore, to compare the expected loss for two scenarios we would need to look at the differences in the rest of the regions.

For $\boldsymbol{x} \in \textcircled{4}$ and $\boldsymbol{x} \in \textcircled{5}$ and biased responses, the expected loss would be the same as the non-strategic case. For $\boldsymbol{x} \in \textcircled{4}$ and $\boldsymbol{x} \in \textcircled{5}$ and the non-biased case, it could be higher or lower. For $\boldsymbol{x} \in \textcircled{3}$, the firm will have a lower (resp. higher) expected loss in the biased responses scenario if the truly unqualified agents are (resp. not) more than truly qualified agents. We furthermore focus on a subset of the region $\textcircled{3}$ identified by Proposition 1, region $\textcircled{3a}$, which is the biased agents that will pass the threshold despite being biased. If we define the region identified by Proposition 1 by $\mathcal{H}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}})$, then region $\textcircled{3a}$ will be $\mathbb{A}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}) \cap \mathbb{A}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}), \theta_{0,\text{NB}}) \cap \mathcal{H}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}})$.

For a setting where the loss function rewards true positives and penalizes false positives as $-u^+TP + u^-FP$, as higher loss is worse as we defined, we can write the following:

$$\begin{aligned}
\mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) &= \boldsymbol{L}_{\textcircled{1}\cup\textcircled{2}}+ \\
&\int_{\boldsymbol{x}\in\textcircled{3}\cup\textcircled{4}\cup\textcircled{5}}\big(-u^+p(\hat{y}=1|\boldsymbol{x},y)f_1(\boldsymbol{x})\alpha_1+u^-p(\hat{y}=1|\boldsymbol{x},y)f_0(\boldsymbol{x})\alpha_0\big)d\boldsymbol{x} \qquad (6)\\
\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) &= \boldsymbol{L}_{\textcircled{1}\cup\textcircled{2}}+ \\
&\int_{\boldsymbol{x}\in\textcircled{3a}}\big(-u^+p(\hat{y}=1|\boldsymbol{x},y)f_1(\boldsymbol{x})\alpha_1+u^-p(\hat{y}=1|\boldsymbol{x},y)f_0(\boldsymbol{x})\alpha_0\big)d\boldsymbol{x} \qquad (7)
\end{aligned}$$

Where $\boldsymbol{L}_{\textcircled{1}\cup\textcircled{2}}$ is the loss coming from regions $\textcircled{1}$ and $\textcircled{2}$ which is present in both scenarios. For $\mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))$, we know all the agents in $\textcircled{3}\cup\textcircled{4}\cup\textcircled{5}$ will be accepted, i.e., $p(\hat{y}=1|\boldsymbol{x}\in\textcircled{3}\cup\textcircled{4}\cup\textcircled{5},y)=1$. Similar for $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))$ and $\boldsymbol{x}\in\textcircled{3a}$.

We can see from equation 6 and equation 7 that depending on the density of label 0 and label 1 agents in the region $\textcircled{3a}$ and comparing it to the region $\textcircled{3}\cup\textcircled{4}\cup\textcircled{5}$ we can have both $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) \leq \mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))$ and $\mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) \leq \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))$ occur. The difference in expected loss lies in the region $\textcircled{3}\cup\textcircled{4}\cup\textcircled{5}-\textcircled{3a}$, or equivalently $\mathbb{S}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}) := \mathbb{A}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})/(\mathbb{A}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})\cap\mathbb{A}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),\theta_{0,\text{NB}})\cap\mathcal{H}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))$, we can write the following for $\mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) - \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) \leq 0$ (resp. ≥ 0):

$$\int_{\boldsymbol{x}\in\mathbb{S}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})}(-u^+f_1(\boldsymbol{x})\alpha_1+u^-f_0(\boldsymbol{x})\alpha_0)dx \leq 0 \text{ (resp. } \geq 0) \tag{8}$$

Therefore, if the density of unqualified agents is higher (resp. lower) than the density of qualified agents over the region $\mathbb{A}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})/(\mathbb{A}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})\cap\mathbb{A}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),\theta_{0,\text{NB}})\cap\mathcal{H}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))$, then:

$$\begin{aligned}
&\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) \leq \mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))\\
&(\text{resp. } \mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) \leq \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})))
\end{aligned}$$

To show the last statement of the proposition, we need to compare $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}),(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))$ and $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{B}}),(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})))$ directly. The difference between these two losses comes from the region where agents will be accepted by $(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})$ and not by $(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})$, and vice versa, after agents' response. Mathematically, for agents responding to $(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})$ without bias, we can show the agents accepted by $(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})$ by $\mathbb{Y}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})\cup\mathbb{A}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})$. We want the intersection of this set with the agents not accepted by $(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})$, which brings us to $\mathbb{T}_1 = (\mathbb{Y}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})\cup\mathbb{A}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))\cap\mathbb{N}(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})$.

Similarly, for agents responding to $(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})$ with bias, we can show the agents accepted by $(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})$ and not by $(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})$ by $(\mathbb{Y}(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})\cap\mathbb{N}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))/\mathbb{A}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})$. However, in this scenario, we need to also account for agents that make it past the actual decision boundary despite being behavioral, i.e., agents in the region $\mathcal{H}(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})\cap\mathbb{A}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{B}}),\theta_{0,\text{B}})$, bringing us to $\mathbb{T}_2 = (\mathcal{H}(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})\cap\mathbb{A}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{B}}),\theta_{0,\text{B}}))\cup((\mathbb{Y}(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}})\cap\mathbb{N}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))/\mathbb{A}(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}}))$.

We need the total loss from region $\mathbb{T}_1$ to be lower than the total loss from the region $\mathbb{T}_2$ in the two scenarios for $\mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) \leq \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{B}}),(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}}))$ to be true. Meaning that we need $\int_{\boldsymbol{x}\in\mathbb{T}_1}(-u^+f_1(\boldsymbol{x})\alpha_1+u^-f_0(\boldsymbol{x})\alpha_0)d\boldsymbol{x} \leq \int_{\boldsymbol{x}\in\mathbb{T}_2}(-u^+f_1(\boldsymbol{x})\alpha_1+u^-f_0(\boldsymbol{x})\alpha_0)d\boldsymbol{x}$ to be true for $\mathbb{L}(\boldsymbol{\theta}_{\text{NB}},(\boldsymbol{\theta}_{\text{NB}},\theta_{0,\text{NB}})) \leq \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{B}}),(\boldsymbol{\theta}_{\text{B}},\theta_{0,\text{B}}))$, and the last inequality of the statement comes from the optimality condition.

B Details of Numerical Experiments

Details for Example 1. For the scenario where the firm is negatively affected by the biased response is Example 1 we used $\boldsymbol{\mu}_1^T = (2,4)$ and $\boldsymbol{\mu}_0^T = (2,3)$ with $\Sigma_1 = \left(\begin{smallmatrix}0.5 & 0\\ 0 & 0.5\end{smallmatrix}\right)$ and $\Sigma_0 = \left(\begin{smallmatrix}1 & 0.5\\ 0.5 & 1\end{smallmatrix}\right)$, and we multiplied the generated data by 10. For the scenario where the firm benefits from agents' biased response we let $\boldsymbol{\mu}_1^T = (3,5)$ and let the rest of the parameters be the same as the first scenario, i.e., $\boldsymbol{\mu}_0^T = (2,3)$ with $\Sigma_1 = \left(\begin{smallmatrix}0.5 & 0\\ 0 & 0.5\end{smallmatrix}\right)$ and $\Sigma_0 = \left(\begin{smallmatrix}1 & 0.5\\ 0.5 & 1\end{smallmatrix}\right)$, and we multiplied the generated data by 10. In both scenarios, we let $B = 5$.

We used the Prelec function described in Section 2 for the behavioral response. Solving the optimization problem takes a considerable amount of time for a large number of data points, here $20,000$, so we used the equivalent of the optimization problem for agents' movement and dictated the movement straight to each data point instead of solving the optimization.

To model agents' behavioral responses, we first identified the agents that would attempt to manipulate their features. Then, we used the movement function with the specified mode, either "B" or "NB", to move the data points and create a new dataset for post-response.

Details for Figure 1 We generated 150 data points using different distributions for each feature. Feature 1 was sampled from $\mathcal{N}(700, 200) - \mathcal{D}((0, 20, 50, 100), (0.6, 0.2, 0.1, 0.1))$ where the second term is a discrete distribution selecting 0 with $p = 0.6$, 20 with $p = 0.2$, 50 with 0.1, and 100 with $p = 0.1$. Feature 2 was sampled from $1500 - \Gamma(4, 100)$. We used a $Score$ column to label each individual for later. The score was calculated from the feature weights $(0.65, 0.35)$. We then used a sigmoid function to assign approval probability and label the sampled data points: $\frac{1}{1+\exp\left(-0.8\times(\frac{x}{10}-80)\right)}$. We assigned the labels using the calculated approval probability and a random number generator. After generating the dataset, we used two copies, one for behavioral response and one for non-behavioral response.

For agents' response to the algorithm, we calculated the agents that can afford the response with a budget of $B = 100$ and performed an optimization problem on only those agents. We convert our model in Section 3 to solve a cost minimization problem for each agent: $\arg\min_{\boldsymbol{x}} cost = \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$ s.t. $\boldsymbol{\theta}^T \boldsymbol{x} \geq \theta_0$. For the behavioral case, we used $\gamma = 0.5$, and the optimization problem $\arg\min_{\boldsymbol{x}} cost = \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$ s.t. $\boldsymbol{w}(\boldsymbol{\theta})^T \boldsymbol{x} \geq \theta_0$.