The Double-Edged Sword of Behavioral Responses in Strategic Classification: Theory and User Studies

Anonymous Author
Anonymous Institution

Abstract

When humans are subject to an algorithmic decision system, they can 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 strategic classification model that considers *behavioral biases* in human responses to algorithms. We show how misperceptions of a classifier (specifically, of its feature weights) can lead to different types of discrepancies between biased and rational agents' responses, and identify when behavioral agents over- or under-invest 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. We complement our analytical results with user studies, which support our hypothesis of behavioral biases in human responses to the algorithm. Together, our findings highlight the need to account for human (cognitive) biases when designing AI systems, and providing explanations of them, to strategic human in the loop.

1 Introduction

As machine learning systems become more widely deployed, including 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 rules and procedures 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, 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 (i.e., 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 to them when the firm offers explanations through an Explainable AI (XAI) method which provides information about feature importance/contribution in the algorithm (e.g. SHAP (Lundberg and Lee, 2017) or LIME (Ribeiro et al., 2016)). We use this model to identify different forms of discrepancies that can arise between behavioral and fully rational agents' responses (Lemmas 1-3). We further identify conditions under which agents' behavioral biases lead them to over- or under-invest in specific features (Proposition 1). 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 2). While the former may be intuitively expected (behaviorally biased agents are less adept at gaming algorithms), the latter is more surprising; we further provide an intu-

Preliminary work. Under review by AISTATS 2025. Do not distribute.

itive explanation for this through a numerical example (Example 1), highlighting the impact of agents' qualification states in determining the ultimate impact of agents' behavioral biases on the firm.

Finally, by conducting a user study, we show that this type of behavioral bias is present when individuals interact with an AI decision assistant. Our study shows that individuals tend to underestimate the importance of the most crucial feature while overestimating the importance of the least important one. We also find that increasing the complexity of the model, either by adding more features or having unbalanced feature weights, amplifies this bias. Additionally, we observe other forms of cognitive biases (not captured by probability weighting biases), such as some individuals disproportionately investing in a feature with a lower starting point when feature weights are similar.

Together, our theoretical findings and user studies highlight the necessity of accounting for not just strategic responses but also cognitive biases when designing AI systems with human in the loop.

Summary of contributions.

- We extend existing models of strategic classification to account for agents' cognitive biases in perceiving feature importance.
- We analyze how these biases lead to over- or under-investment in certain features compared to fully rational agents. We further show that behaviorally biased agents can increase or decrease firm utility.
- Through a user study, we confirm that cognitive biases influence human's understanding of and responses to AI systems, especially when they are given (explanations of) models with unbalanced feature weights and a higher number of features.

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; Harris et al., 2022) or partial information (Haghtalab et al., 2023; Cohen et al., 2024) about the algorithm. While our base model of agents' strategic responses to (threshold) 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 when responding to algorithmic decision rules (other than classification) has been considered in recent work (Morewedge et al., 2023; Zhu et al., 2024; Liu et al., 2024; Heidari et al., 2021; Ethayarajh et al., 2024). 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 machine learning. 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., 2019; 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. We further confirm the presence of behavioral biases through a user study. Previous works have utilized user studies to assess interpretable models based on factors such as time spent, number of words, and accuracy (Lakkaraju et al., 2016), to establish the core principles of interpretability goals Hong et al. (2020), and to assess the impact of model interpretability on predicting model outputs (Poursabzi-Sangdeh et al., 2021). In contrast, we assess how behavioral biases can result in human subjects' sub-optimal responses to interpretable models. We review additional related work in Appendix A.

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 $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_n]^T$ denotes the *feature weights*; we assume features 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:

$$\begin{aligned} \boldsymbol{x}_{\text{NB}} := \arg\max_{\boldsymbol{x}} \; & r h(\boldsymbol{x}, (\boldsymbol{\theta}, \theta_0)) - c(\boldsymbol{x}, \boldsymbol{x}_0) \\ \text{subject to} \quad & c(\boldsymbol{x}, \boldsymbol{x}_0) \leq B \end{aligned} \tag{1}$$

where $r > 0$ is the reward of positive classification, $c(\boldsymbol{x}, \boldsymbol{x}_0)$ is the cost of changing feature vector $\boldsymbol{x}_0$ to $\boldsymbol{x}$, and B is the agent's budget. We consider three different cost functions: *norm-2 cost* (with $c(\boldsymbol{x}, \boldsymbol{x}_0) = \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2 = \sum_i (x_i - x_{i,0})^2$), *quadratic cost* (with $c(\boldsymbol{x}, \boldsymbol{x}_0) = \sum_i c_i (x_i - x_{0,i})^2$), and *weighted Manhattan (taxicab) distance cost* (with $c(\boldsymbol{x}, \boldsymbol{x}_0) = \boldsymbol{c}^T |\boldsymbol{x} - \boldsymbol{x}_0| = \sum_i c_i (|x_i - x_{0,i}|)$). Our analytical results are presented for the *norm-2 cost*. We also characterize the agent's best-responses under other cost functions to highlight that similar agent behavior can be seen under them.

Anticipating the agents' responses, the firm can 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., weighted sum of TP and FP costs).

Behavioral Responses. We extend the strategic classification model to allow for behavioral responses by agents. Formally, recall that we normalize the feature weight vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_n]^T$ to have $\sum_i \theta_i = 1$. We interpret it as a probability vector, and assume that behaviorally biased agents misperceive $\boldsymbol{\theta}$ as $\boldsymbol{w}(\boldsymbol{\theta})$, where $\boldsymbol{w}(\cdot)$ is a function capturing their biases. One choice for $\boldsymbol{w}(\cdot)$ can be $w_j(\boldsymbol{\theta}) = p(\sum_{i=1}^{j} \theta_i) - p(\sum_{i=1}^{j-1} \theta_i)$ (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:

$$\begin{aligned} \boldsymbol{x}_{\text{B}} := \arg\max_{\boldsymbol{x}} \; & rh(\boldsymbol{x}, (\boldsymbol{w}(\boldsymbol{\theta}), \theta_0)) - c(\boldsymbol{x}, \boldsymbol{x}_0) \\ \text{subject to} \quad & c(\boldsymbol{x}, \boldsymbol{x}_0) \le B \end{aligned} \tag{2}$$

Note that the agent now responds to a *perceived feature weights* $(\boldsymbol{w}(\boldsymbol{\theta}), \theta_0)$. In return, while always accounting for agents' strategic behavior ("gaming"), we assume the firm may or may not be aware that agents have behavioral biases when gaming the system. 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 (potentially different) classifier $(\boldsymbol{\theta}', \theta_0)$. Then, if a firm is aware of strategic agents' behavioral biases, it selects the threshold $(\boldsymbol{\theta}_{\text{B}}, \theta_{0,\text{B}}) := \arg\min_{(\boldsymbol{\theta},\theta_0)} \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}), (\boldsymbol{\theta}, \theta_0))$ and incurs a loss $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_B), (\boldsymbol{\theta}_B, \theta_{0,\text{B}}))$. On the other hand, a firm that assumes agents are fully rational selects the threshold classifier $(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}})$, yet incurs the loss $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}), (\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}))$.

3 Agents' Strategic Responses

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 $\boldsymbol{x}_{\text{NB}}$ (the solution to equation 1) and $\boldsymbol{x}_{\text{B}}$ (the solution to equation 2) 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 $0 < d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0) \le B$,*

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

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 1 and 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 1 believe they are accepted without effort, while those in 6 believe they do not have sufficient budget to succeed. Agents in regions 2 and 5 manipulate their features unnecessarily (they would not, had they been fully rational), and again, for different reasons: agents in 2 are not accepted even at their highest effort level, while those in 5 believe they must reach the boundary but they would be accepted regardless of their effort. Finally, in region 3, agents *undershoot* the actual boundary (i.e., exert less effort than needed due to their biases), while those in region 4 *overshoot* (i.e., exert more effort than needed to get accepted).

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

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

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

Figure 1: (a) Fully rational and (b) Biased responses, and (c) Classes of differing actions under quadratic costs.

(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 who perceive the decision boundary to be closer to them than it truly is (regions 2 and 3 in Figure 1c) will underinvest in the features for which they underestimate the importance. Similarly, agents that perceive the boundary to be farther (regions 4 and 5 in Figure 1c) will over-invest in the features for which they overestimate the importance.

3.1 Alternative Cost Functions

We next show that regions of differing responses between behavioral and non-behavioral agents, similar to those depicted in Figure 1c, will also emerge under quadratic and weighted Manhattan cost functions.

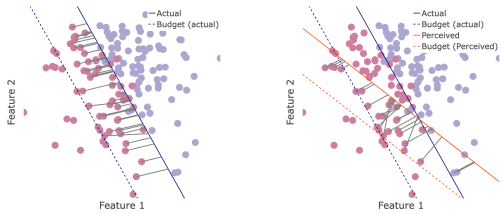
Quadratic Cost Function. The following lemma characterizes the post-strategic features under the quadratic cost $c(\boldsymbol{x}, \boldsymbol{x}_0) = \sum_i c_i (x_i - x_{i,0})^2$.

Lemma 2. *Let* $\boldsymbol{C}$ *denote a diagonal matrix with* c_i*'s as its diagonal, and let* $\boldsymbol{y}$ *denote the feature vectors satisfying* $\boldsymbol{\theta}^T \boldsymbol{y} = \theta_0$. *For an agent with starting feature vector* $\boldsymbol{x}_0$, *if* $\boldsymbol{x}_0$ *is in the n-dimensional ellipsoid described by* B *and* $\boldsymbol{C}$, *i.e., if* $(\boldsymbol{y} - \boldsymbol{x}_0)^T \boldsymbol{C} (\boldsymbol{y} - \boldsymbol{x}_0) \leq B$,

$$x_{NB,i} = \frac{\theta_0 - \boldsymbol{\theta}^T \boldsymbol{x}_0}{\sum_j \frac{\theta_j^2}{c_j}} \cdot \frac{\theta_i}{c_i} + x_{0,i} \ .$$

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 2 illustrates the best-responses of Lemma 2 for rational (non-behavioral) and biased (behavioral) agents. Specifically, the condition in Lemma 2 constructs an n-dimensional ellipsoid around every point on the line $\boldsymbol{\theta}^T \boldsymbol{x} = \theta_0$, containing agents who have sufficient budget to strategically change their features to reach that particular point on the boundary, with the coefficients c_i determining the scaling along each axis. Since the scaling of the ellipsoid does not depend on the point of the line we are focusing on, the union of these ellipsoids (determining the set of all agents who can afford to be classified positively through gaming) forms a band below the line $\boldsymbol{\theta}^T \boldsymbol{x} = \theta_0$. Note that this band differs from the one in Lemma 1.



(a) Rational response (b) Biased response

Figure 2: Strategic responses under quadratic costs.

Weighted Manhattan Distance Cost Function. The following lemma characterizes the post-strategic features under the weighted Manhattan cost $c(\boldsymbol{x}, \boldsymbol{x}_0) = \sum_i c_i |x_i - x_{i,0}|$.

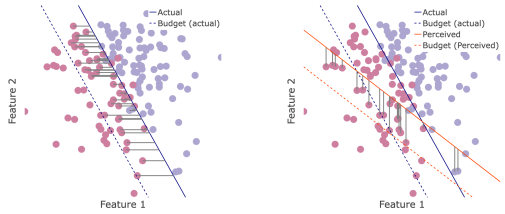
Lemma 3. *Let* $\boldsymbol{e}_i$ *be the unit vector with 1 in the* i^{th} *coordinate and 0 elsewhere, and* $k = \arg\min_i \frac{c_i}{\theta_i}$. *For an agent with starting feature* $\boldsymbol{x}_0$, *if* $\boldsymbol{\theta}^T \boldsymbol{x}_0 + \frac{\theta_k}{c_k} B \geq \theta_0$,

$$\boldsymbol{x}_{NB} = \boldsymbol{x}_0 + (\theta_0 - \boldsymbol{\theta}^T \boldsymbol{x}_0) \frac{c_k}{\theta_k} \boldsymbol{e}_k \ .$$

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 3 illustrates the best-responses of Lemma 3 for rational (non-behavioral) and biased (behavioral) agents. Again, the set of agents who can afford to game the system to receive a positive classification, as identified in Lemma 3, is a band below the classifier (but different from those of Lemmas 1 and 2). In particular, under this cost, agents spend all their budget on changing the feature with the most "bang-for-the-buck" $\frac{c_i}{\theta_i}$ (or perceived bang-for-the-buck $\frac{c_i}{w_i(\boldsymbol{\theta})}$). As seen in the two-dimensional illustration in Figure 3, this means that while it is optimal for rational agents

to invest only in feature 2, those with behavioral bias believe feature 1 has a better return, leading to a suboptimal response by them. We also note that even though the movements of agents in the specified band are different from the movement for the norm-2 cost, the bands form the same regions of differing responses as in Figure 1, where agents overshoot, undershoot, do nothing at all, or needlessly change their features, when they are behaviorally biased.



(a) Rational response (b) Biased response

Figure 3: Strategic responses under Manhattan costs.

Cost Function in the User Studies. For our human subject experiments in Section 5, we describe the cost of changing features to participants through a *piecewise linear cost function*. This can be viewed as an approximation of a quadratic cost using a step function with a weighted Manhattan cost at each step, with the approximation improving as the number of steps increases (see the two-dimensional illustration in Figure 4). Specifically, in our user experiments, we break the budget B into three steps of increments B_1, B_2, and B_3 with $B_1 + B_2 + B_3 = B$, and assign a constant cost c_1, c_2, and c_3 for changing features at each increment. This means that in each step, agents face a weighted Manhattan cost, but overall, the cost is not fixed, and investing in a single feature is not optimal.

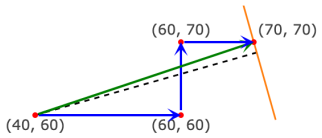


Figure 4: Strategic responses under a quadratic cost (green) vs. a piece-wise linear cost function (blue).

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 and agents' welfare. 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 demonstrated in Section 3, behavioral agents may overshoot or undershoot the threshold when gaming the algorithm (compared to rational agents); 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 decrease the firm's utility. In other words, perhaps unexpectedly, in these situations, the firm would prefer rational agents, who are better at gaming the system, to behaviorally biased agents, who are worse at gaming the system. The following example numerically illustrates this.

Example 1. *Consider a setting where we have a 2D 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 three scenarios; the first two scenarios only differ in the mean $\boldsymbol{\mu}_1$ choice, and the third scenario differs with these in $\boldsymbol{\mu}_1$, Σ_1, Σ_0, and B (see Appendix C for details). The first two scenarios (top and middle rows in Figure 5) are baselines: we consider an* oblivious *firm that chooses its classifier without accounting for any strategic response (whether rational or behavioral) from agents. This helps us hone in on the impacts of agents' qualification states on the firm's utility. Then, in the third scenario (bottom row in Figure 5), we consider a firm that is aware of strategic behavior (and any behavioral biases) by agents and optimally adjusts its classifier. For each scenario, Figure 5 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). The firm's utility in each case is shown at the top of the corresponding subplot.*

We start with the baselines (an oblivious firm that keeps the classifier fixed). In the top row scenario, the firm is negatively impacted by agents' behavioral biases, while in the middle row scenario, the firm benefits from agents' biases (both compared to the fully rational setting). 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 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.

Next, we consider the (non-oblivious) firm that adjusts its classifier optimally (accounting for strategic responses and behavioral biases, if any). We observe that even though the firm is aware of agents' bias, its loss is higher than the case of rational responses. As seen in the left panel of the bottom row of Figure 5, more regions can impact the loss than in Figure 1. The

most important regions in this scenario are the areas accepted by $\boldsymbol{\theta}_{NB}$ but not by $\boldsymbol{\theta}_B$ (after response), and vice versa. As there are more qualified than unqualified agents in these two regions, the firm is negatively impacted by agents' bias (compared to fully rational agents) even though the firm is aware of the bias.

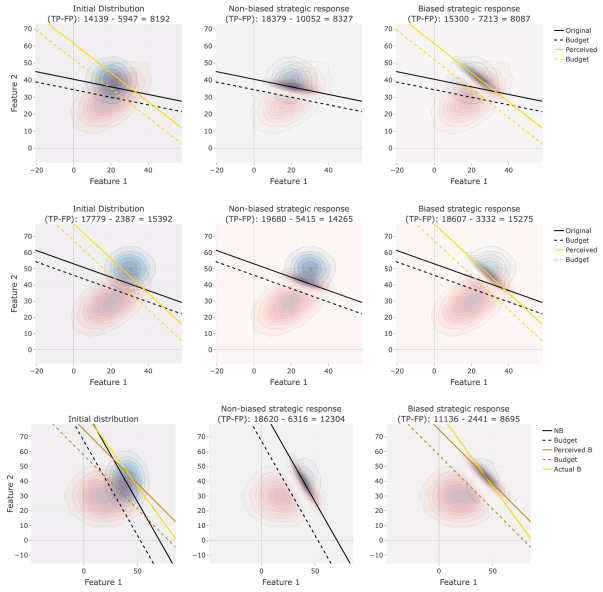


Figure 5: An oblivious firm may have lower (top) or higher (middle) utility when agents are biased (vs. rational). A non-oblivious firm may still have a lower utility when agents are biased (bottom).

The next proposition formalizes the above intuition.

Proposition 2. *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:*

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

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

$$\max\{\mathbb{L}(\boldsymbol{\theta}_{NB}, (\boldsymbol{\theta}_{NB}, \theta_{0,NB})), \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_B), (\boldsymbol{\theta}_B, \theta_{0,B}))\} \leq \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{NB}), (\boldsymbol{\theta}_{NB}, \theta_{0,NB})) \quad (4)$$

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

$$\mathbb{L}(\boldsymbol{\theta}_{NB}, (\boldsymbol{\theta}_{NB}, \theta_{0,NB})) \leq \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_B), (\boldsymbol{\theta}_B, \theta_{0,B})) \leq \mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{NB}), (\boldsymbol{\theta}_{NB}, \theta_{0,NB})) \quad (5)$$

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.

Agents' Welfare: We end this section by comparing the impacts of behavioral biases on agents' welfare (sum of their utilities). As a baseline, note that if the firm was oblivious to agents' strategic responses and did not adjust the classifier, agents would have lower welfare when they are behaviorally biased (compared to when rational). This is an expected outcome since behaviorally biased agents are worse at gaming the algorithm and respond sub-optimally. But, perhaps more unexpectedly, when the firm adjusts its classifier in response to agents' strategic behavior and behavioral biases, various scenarios can occur. For instance, in the bottom row of Figure 5, qualified agents have *higher* social welfare when they are behaviorally biased compared to if they had been rational. We provide additional details on reasons for this in Appendix D.

5 User Study

5.1 Study Design, Participants, and Setup

To understand how human biases affect their responses to algorithms, we conducted a large-scale online survey where participants completed a task with an optimal solution. We then measured how their answers differed from the optimal solution to assess bias. Similar

[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.

to previous sections, we focus on how participants perceive the feature weights and how this influences their response to the algorithm. Surveys generally took 5.5 minutes, and participants were paid $16 per hour. The study design was reviewed by our IRB, and the full protocol is included in the supplementary materials.

Experimental Design. We used a 2×2 between-subjects design, varying the number of features and the feature weights. Participants were randomly assigned to a condition and shown a single explanation (shown in Figure 6), with a) either two or four features and b) either balanced or unbalanced feature weights.

Procedure. All participants were shown a truthful and complete explanation from an interpretable ML algorithm. Each participant was asked to complete a task based on the information provided in the explanation. After completing the task, participants were asked questions about their understanding, trust, satisfaction, and performance, common measures in explainability evaluations (Mohseni et al., 2021).

Explanation. Based on prior work, which has found that feature importance is the most used explanation (Nauta et al., 2023) and that visualization can help users understand explanations (Adadi and Berrada, 2018), we developed an explanation that shows feature weights in a bar graph (Figure 6). The system uses the displayed feature weights as an interpretable ML algorithm.

Task. The task asked participants to give advice to a family member about how to prepare for the college application process. Participants were told an AI system provides predictions of college admissions. The features were resume (R) and cover letter (CL), with two additional features, interview (I) and LinkedIn profile (LP), for the four feature conditions.

Each participant had a budget (10 hours) to allocate between the given features to improve the likelihood of acceptance recommendation. For starting features, in the two-feature and four-feature scenarios, we used $\boldsymbol{x}_0 = (40\ (\text{R})\ , 60\ (\text{CL}))$ and $\boldsymbol{x}_0 = (60\ (\text{I})\ , 40\ (\text{R})\ , 60\ (\text{CL})\ , 65\ (\text{LP}))$, where, the maximum for feature scores is 100. For each hour allocated to any feature, participants were given a piecewise linear cost: the feature's score improves by 5 points for the first four hours, 2.5 points for the second four hours, and 1 point for extra hours after that.

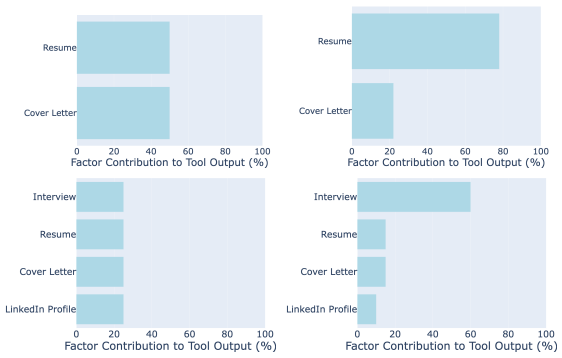


Figure 6: The scenarios shown to participants: two or four features of similar importance (top and bottom left resp.), and two or four features of differing importance (top and bottom right).

Correct ("optimal") answers. For two balanced features, one should *not* invest more than 6 hours in any feature. In the scenario with two unbalanced features, the optimal investments in the resume and cover letter are 8 hours and 2 hours, respectively. For the scenario with four balanced features, the optimal is to invest at most 4 hours in any feature. For the unbalanced four features case, the optimal investment is to allocate 8 hours to the interview and the remaining 2 hours to the resume and cover letter. In this case, any investment in the LinkedIn profile feature is sub-optimal. A more detailed explanation is given in Appendix E.

Measures. For other dependent measures, we used self-reported measurements of satisfaction, understanding, trust, and task performance using five-point semantic scales. We lightly edited questions from Mohseni et al. (2021) for brevity and clarity.

Recruitment. We recruited 100 participants through Prolific in September 2024. Quotas on education level and gender ensured the sample was representative of the United States. Additionally, we gathered demographic information and assessed participants' familiarity with machine learning to ensure a representative and unbiased sample.

5.2 Results and Discussion

Adding complexity reduces performance. Our findings indicate that participant performance decreases when we increase the number of features from two to four or shift from balanced to unbalanced feature weights. We evaluate performance by comparing the total score of a response to the optimal total score for that case. The total score of each response is calculated by first determining the new feature vector $\boldsymbol{x}$ by adding the improvements of each feature to $\boldsymbol{x}_0$ based on the subject's response, and then calculating $\boldsymbol{\theta}^T\boldsymbol{x}$.

In Figure 7, we observe that as the number of features increases, adding more complexity to the model, participants will move further away from the optimal score in balanced and unbalanced cases. The increase in the unbalanced case is almost double that of the balanced case. This indicates that we observe behavioral

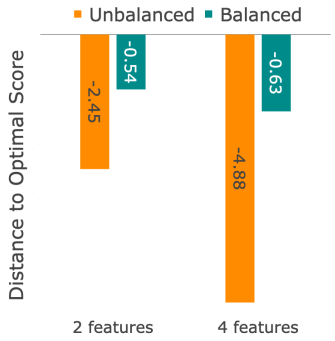


Figure 7: The average distance to optimal ($\boldsymbol{\theta}^T \boldsymbol{x}_{\text{NB}} - \boldsymbol{\theta}^T \boldsymbol{x}_{\text{B}}$) for the four scenarios.

responses even in the balanced case, indicating biases beyond those in our theoretical predictions. For a fixed number of features, we see that answers are far from optimal when we have unbalanced features. Increasing the number of features will also lead to worse answers.

Table 1: Number of responses in each scenario, (B) balanced and (U) unbalanced features

Scenario	Opt.	1-feature	Sub-opt.
2-features (B)	21	0	6
4-features (B)	14	1	10
2-features (U)	5	3	16
4-features (U)	1	1	22

Table 1 shows that not only does the average distance from the optimal score increase with added complexity, but also the number of participants that responded sub-optimally increases. Most participants could find the optimal answers in the balanced cases, but most responded sub-optimally in the unbalanced cases.

Participants exhibit different behavioral biases. The unbalanced scenarios indicate that most participants' behavior is consistent with following a Prelec function when (mis)perceiving feature importance. This leads them to under-invest in the most important feature and over-invest in the least important one. Participants not following the Prelec function tend to allocate all their budget to the most important feature.

The results from the balanced scenarios shed light on another behavioral bias: In the case of similar importance, participants invest more in the feature with a lower starting point (the resume). In the balanced scenarios, we notice that most participants respond without a behavioral bias, as predicted by the bias and Prelec functions. However, some participants responded sub-optimally, all over-investing in the resume. In the unbalanced four-feature case, the average investment in the resume is higher despite the resume and the cover letter having the same importance, indicating that this occurs for any two features with the same weight regardless of the importance of other features.

Focusing on the unbalanced scenarios, we see that the number of participants who respond with the optimal answer drops when we increase the number of features. More participants decide to invest all their budget in the most important feature. Findings from the unbalanced two-feature scenario show that if participants do not invest all their budget in the most important feature, they under-invest in it.

Table 2: Average distance of investment in most important and least important features in unbalanced scenarios from the optimal.

	Most important	Least important
2-features	-2.13 hours	+2.13 hours
4-features	-4.11 hours	+1.76 hours

Assuming the Prelec function, we find that γ for the participants answering the unbalanced two-features scenario is $\gamma \leq 0.64$, vs. $\gamma \leq 0.55$ for four-features. (The lower the γ in the Prelec function, the more intense the bias.) These upper bounds come from the fact that participants must underestimate the importance of the most important feature enough so that they conclude it is better to invest in the second most important feature.

Another interesting observation in the unbalanced four-features scenario is that, even though any investment in the least important feature was sub-optimal, 18 participants still invested. This could be either a result of participants overestimating the importance of the least important feature, or a different behavioral bias, where participants prefer to invest in all options, and avoid leaving any feature as is.

6 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 and agents' welfare. Furthermore, through a user study, we support our theoretical model and results, showing that most participants respond sub-optimally when provided with an explanation of feature importance/contribution. Exploring analytical models accounting for biases beyond misperception of feature weights, and exploring the possibility of designing explanation methods that can help mitigate biases, remain as important directions for further investigation.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

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)*.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805.

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.

Camacho, A. and Conover, E. (2011). Manipulation of social program eligibility. *American Economic Journal: Economic Policy*, 3(2):41–65.

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., Weller, A., and Wood, A. (2019). Accountability of ai under the law: The role of explanation.

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.

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *SIGKDD Explor. Newsl.*, 15(1):1–10.

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. (2023). Calibrated stackelberg games: Learning optimal commitments against calibrated agents.

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.

Hong, S. R., Hullman, J., and Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and needs. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

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.

Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. (2022). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, 55(5).

Karimi, A.-H., Schölkopf, B., and Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 353–362, New York, NY, USA. Association for Computing Machinery.

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

Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, page 126–137, New York, NY, USA. Association for Computing Machinery.

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., Bach, S. H., and Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1675–1684, New York, NY, USA. Association for Computing Machinery.

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.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.

Mohseni, S., Zarei, N., and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Trans. Interact. Intell. Syst.*, 11(3–4).

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.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55(13s).

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.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

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

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Sixt, L., Schuessler, M., Popescu, O.-I., Weiß, P., and Landgraf, T. (2022). Do users benefit from interpretable vision? a user study, baseline, and dataset. In *International Conference on Learning Representations*.

Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 10–19, New York, NY, USA. Association for Computing Machinery.

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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   The model and assumptions are described in Section 2. Additional assumptions for propositions, if any, are mentioned in the body of each proposition.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
   The anonymized source code is available in the supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   The main assumptions on our proposed model are described in Section 2. Additional assumptions for the presented lemmas and propositions, if any, are mentioned in the body of each proposition.
   (b) Complete proofs of all theoretical results. [Yes]
   All proofs for theoretical results are available in Appendix B.
   (c) Clear explanations of any assumptions. [Yes]
   Explanations are given in the main text, in footnotes, or through intuitive explanations proceeding theoretical results.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   All the details for numerical experiments are available in the Appendix C. The anonymized source code is available in the supplementary material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes]
   The full text of the survey given to participants is available in the supplementary materials. Screenshots were omitted for anonymity.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes]
   The consent form and IRB approval are available but were not included in the supplementary material to preserve anonymity. If requested, we can prepare and share anonymized versions of these documents with the reviewers. We will include these in the final submission and provide them upon request. The IRB approval number is 810577.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes]
   This is included in the main body of the paper in Section 5.