Supplementary Materials

A Additional related work

Iterpretable Machine Learning and Explainable AI: The interpretability of machine learning models and explainable AI is receiving more attention as it becomes more necessary for firms in various fields of work to explain an AI decision-making assistant or understand the algorithm's process for that specific decision (Ali et al., 2023; Adadi and Berrada, 2018). Many studies have focused on providing guidelines and new objectives for the algorithm to ensure interpretability, Freitas (2014) discusses the interpretability issues for five specific classification models and more generic issues of interpretability. Lakkaraju et al. (2016) provides a multi-objective optimization problem and uses model interpretability as a goal of the learning algorithm, Lundberg and Lee (2017) and Ribeiro et al. (2016) provide methods for posthoc explanations. Adebayo et al. (2018) provides a method to evaluate an explanation method for image data. Many other works have also conducted user studies for finding or evaluating the guidelines using measures such as satisfaction, understanding, trust, etc. (Poursabzi-Sangdeh et al., 2021; Kulesza et al., 2015; Sixt et al., 2022).

Strategic Classification: Explanations enable users to potentially respond (Camacho and Conover, 2011) to an algorithm to improve, meaning that they change their features to change their actual qualification or cheat, meaning that they manipulate their features to game the algorithm. This topic is extensively discussed in works such as Perdomo et al. (2020); Hardt et al. (2016); Liu et al. (2020); Bechavod et al. (2021); Hu et al. (2019). However, these works assume complete information on the model parameters, which is not necessarily a correct assumption. Cohen et al. (2024) explores the partial information released by the firm and discusses the firm's optimization problem and agents' best response. Haghtalab et al. (2023) introduced the calibrated Stackelberg games where the agent does not have direct access to the firm's action. This can also be implemented in our framework where the firm uses $\boldsymbol{\theta}$ but announces $\boldsymbol{\theta}'$, and agents respond to $\boldsymbol{w}(\boldsymbol{\theta}')$. Another line of work called actionable recourse suggests giving actionable responses to users alongside the model explanation could be beneficial and help the users have a better outcome (Karimi et al., 2022). Ustun et al. (2019) provides an integer programming toolkit and introduces actionable recourse in linear regression. Karimi et al. (2021) introduces algorithmic recourse that allows people to act rather than understand. Harris et al. (2022) combines the algorithmic recourse with partial information and has a firm that provides actionable recourse and steers agents. They show that agents and the firm are never worse off in expectation in this setting.

B Proofs

Proof of Lemma 1, Lemma 2, and Lemma 3 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:

$$\begin{aligned} \boldsymbol{x}_{\text{NB}} := \arg\max_{\boldsymbol{x}} \;\; & r - c(\boldsymbol{x}, \boldsymbol{x}_0) \\ \text{subject to} \quad & \boldsymbol{\theta}^T \boldsymbol{x} = \theta_0 \end{aligned} \tag{6}$$

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 $\boldsymbol{x}_{\text{NB}} = \boldsymbol{x}_0$.

Lemma 1: 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 $\boldsymbol{x}_{\text{NB}} = \boldsymbol{x}_0 + d(\boldsymbol{x}_0, \boldsymbol{\theta}, \theta_0)\boldsymbol{\theta}$.

Lemma 2 The quadratic cost is similar to norm-2 cost, by directly solving the optimization problem and having λ to be the Lagrange multiplier for the constraint we find:

$$x_{i,\text{NB}} = \frac{\lambda}{2}\frac{\theta_i}{c_i} + x_{i,0} \text{ and } \frac{\lambda}{2} = \frac{\theta_0 - \boldsymbol{\theta}^T \boldsymbol{x}_0}{\sum_j \frac{\theta_j^2}{c_j}} \Rightarrow x_{i,\text{NB}} = \frac{\theta_0 - \boldsymbol{\theta}^T \boldsymbol{x}_0}{\sum_j \frac{\theta_j^2}{c_i}} \frac{\theta_i}{c_i} + x_{i,0} \tag{7}$$

Which is, in some sense, a movement with a weighted distance from $\boldsymbol{x}_0$ towards the hyperplane.

Lemma 3 For the weighted Manhattan cost we are aiming to find the most efficient feature, i.e., the feature with the lowest $\frac{c_i}{\theta_i}$.

Proof of Proposition 1 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 2 We start the proof from the leftmost inequality in equation 3. 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.

We next provide a characterization of the set of agents who fall within regions 1 and 3 in Figure 1c. These are the set of agents who will still pass the (true) decision boundary regardless of their biases.

Lemma 4. *For a given $(\boldsymbol{\theta}, \theta_0)$, agents that satisfy $(1 - \sigma(\boldsymbol{\theta}))\theta_0 \leq (\boldsymbol{\theta} - \sigma(\boldsymbol{\theta})\boldsymbol{w}(\boldsymbol{\theta}))^T \boldsymbol{x}$, if given enough budget, 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.*

Proof. 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:

$$\boldsymbol{\theta}_0 \leq \frac{\theta_0 - \boldsymbol{w}(\boldsymbol{\theta})^T \boldsymbol{x}}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2} \boldsymbol{\theta}^T \boldsymbol{w}(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \boldsymbol{x} = \frac{\boldsymbol{\theta}^T \boldsymbol{w}(\boldsymbol{\theta})}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2} \theta_0 + \left(\boldsymbol{\theta} - \frac{\boldsymbol{\theta}^T \boldsymbol{w}(\boldsymbol{\theta})}{\|\boldsymbol{w}(\boldsymbol{\theta})\|^2} \boldsymbol{w}(\boldsymbol{\theta}) \right)^T \boldsymbol{x}$$
$$\Rightarrow (1 - \sigma(\boldsymbol{\theta}))\theta_0 \leq (\boldsymbol{\theta} - \sigma(\boldsymbol{\theta})\boldsymbol{w}(\boldsymbol{\theta}))^T \boldsymbol{x} \tag{8}$$

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

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):

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

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

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

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

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 $\boldsymbol{x} \in$ ① and $\boldsymbol{x} \in$ 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 $\boldsymbol{x} \in$ ④ and $\boldsymbol{x} \in$ ⑤ and biased responses, the expected loss would be the same as the non-strategic case. For $\boldsymbol{x} \in$ ④ and $\boldsymbol{x} \in$ ⑤ and the non-biased case, it could be higher or lower. For $\boldsymbol{x} \in$ 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 ③ identified by Lemma 4, region (3*a*), which is the biased agents that will pass the threshold despite being biased. If we define the region identified by Lemma 4 by $\mathcal{H}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}})$, then region (3*a*) 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:

$$\mathbb{L}(\boldsymbol{\theta}_{\text{NB}}, (\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}})) = \boldsymbol{L}_{①\cup②} + \int_{\boldsymbol{x} \in ③\cup④\cup⑤} \big(- u^+ p(\hat{y} = 1 | \boldsymbol{x}, y) f_1(\boldsymbol{x}) \alpha_1 + u^- p(\hat{y} = 1 | \boldsymbol{x}, y) f_0(\boldsymbol{x}) \alpha_0 \big) d\boldsymbol{x} \tag{9}$$

$$\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}), (\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}})) = \boldsymbol{L}_{①\cup②} + \int_{\boldsymbol{x} \in (3a)} \big(- u^+ p(\hat{y} = 1 | \boldsymbol{x}, y) f_1(\boldsymbol{x}) \alpha_1 + u^- p(\hat{y} = 1 | \boldsymbol{x}, y) f_0(\boldsymbol{x}) \alpha_0 \big) d\boldsymbol{x} \tag{10}$$

Where $\boldsymbol{L}_{①\cup②}$ is the loss coming from regions ① and 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 ③∪④∪⑤ will be accepted, i.e., $p(\hat{y} = 1 | \boldsymbol{x} \in ③\cup④\cup⑤, y) = 1$. Similar for $\mathbb{L}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}), (\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}))$ and $\boldsymbol{x} \in$ (3*a*).

We can see from equation 9 and equation 10 that depending on the density of label 0 and label 1 agents in the region (3*a*) and comparing it to the region ③ ∪ ④ ∪ 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 ③ ∪ ④ ∪ ⑤ − (3*a*), or equivalently $\mathbb{S}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}) \coloneqq \mathbb{A}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}) / (\mathbb{A}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}) \cap \mathbb{A}(\boldsymbol{w}(\boldsymbol{\theta}_{\text{NB}}), \theta_{0,\text{NB}}) \cap \mathcal{H}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}}))$, 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):

$$\int_{\boldsymbol{x} \in \mathbb{S}(\boldsymbol{\theta}_{\text{NB}}, \theta_{0,\text{NB}})} (-u^+ f_1(\boldsymbol{x}) \alpha_1 + u^- f_0(\boldsymbol{x}) \alpha_0) dx \leq 0 \text{ (resp. } \geq 0) \tag{11}$$

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:

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

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.

C Details of Numerical Experiments

Details for Example 1 and Figure 5 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.

For the last row of Figure 5 we used $\boldsymbol{\mu}_1^T = (4, 4)$ and $\boldsymbol{\mu}_0^T = (2, 3)$ with $\Sigma_1 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $\Sigma_0 = \left(\begin{smallmatrix} 3 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, and we multiplied the generated data by 10. We used $B = 10$.

Details for Figure 1, Figure 2, and Figure 3 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: $\frac{1}{1+\exp\left(-0.8\times\left(\frac{x}{10}-80\right)\right)}$. 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.

In Figure 1, 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 solved a cost minimization problem for each agent in the band specified by Lemma 1: $\arg\min_{\boldsymbol{x}} cost = \|\boldsymbol{x} - \boldsymbol{x}_0\|_2$ 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. $\boldsymbol{w}(\boldsymbol{\theta})^T\boldsymbol{x} \geq \theta_0$.

In Figure 2, 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 solved a cost minimization problem for each agent in the band specified by Lemma 2: $\arg\min_{\boldsymbol{x}} cost = \sum_i c_i(x_i - x_{0,i})^2$ s.t. $\boldsymbol{\theta}^T\boldsymbol{x} \geq \theta_0$. For the behavioral case, we used $\gamma = 0.5$, and the optimization problem $\arg\min_{\boldsymbol{x}} cost = \sum_i c_i(x_i - x_{0,i})^2$ s.t. $\boldsymbol{w}(\boldsymbol{\theta})^T\boldsymbol{x} \geq \theta_0$.

In Figure 3, 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 solved a cost minimization problem for each agent in the band specified by Lemma 3: $\arg\min_{\boldsymbol{x}} cost = \boldsymbol{c}^T|\boldsymbol{x} - \boldsymbol{x}_0|$ s.t. $\boldsymbol{\theta}^T\boldsymbol{x} \geq \theta_0$. For the behavioral case, we used $\gamma = 0.5$, and the optimization problem $\arg\min_{\boldsymbol{x}} cost = \boldsymbol{c}^T|\boldsymbol{x} - \boldsymbol{x}_0|$ s.t. $\boldsymbol{w}(\boldsymbol{\theta})^T\boldsymbol{x} \geq \theta_0$.

D Agents' Welfare

Figure 8 highlights the change in utility when agents are behaviorally biased (vs. when they were rational) across different regions in the feature space, with the regions generated based on the firm's optimal choice of threshold and agents' responses to it. In particular, the utility of agents in the green-highlighted region (this is $\mathbb{Y}(\boldsymbol{\theta}_{\mathrm{B}}, \theta_{0,\mathrm{B}}) \cap \mathbb{N}(\boldsymbol{\theta}_{\mathrm{NB}}, \theta_{0,\mathrm{NB}})$ in Proposition 2) increases when they are behaviorally biased. One subset of agents in this region are those who in the rational case exert effort to get admitted and have a utility $r - c(\boldsymbol{x}, \boldsymbol{x}_0)$, whereas in the behaviorally biased case they attain utility $r > r - c(\boldsymbol{x}, \boldsymbol{x}_0)$ as they get admitted without any effort (and they correctly assume so). Another one is the subset of agents who would not try to get to the decision boundary in the rational case (and so have utility of 0), but in the behavioral case, they are receiving utility r without any movement and due to the change of the decision boundary. For the numerical example in the bottom row of Figure 5, there are more agents in this green-highlighted region than in the remaining red-highlighted regions (where biased agents have lower utility than rational agents), leading to an overall higher welfare for all agents when they are biased compared to when they were rational.

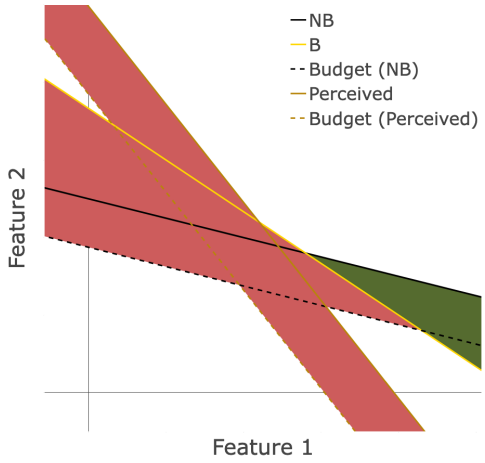

Figure 8: Regions where agents have higher (green) or lower (red) utility when biased vs. when rational.

E Piece-wise Cost Function Solution

Consider a setting similar to the piece-wise cost function described. To decide the feature to spend B_1 of the budget, we are comparing $\frac{c_1}{\theta_1}$, $\frac{c_1}{\theta_2}$, and $\frac{c_1}{\theta_3}$ as they all have the same cost for the first step of the budget. Without loss of generality imagine we have $\frac{c_1}{\theta_1} < \frac{c_1}{\theta_2} < \frac{c_1}{\theta_3}$, therefore, we choose to allocate the B_1 amount of our budget to the first feature. For B_2, we do a similar comparison but we have to use c_2 for the first feature since the first feature is now in the second step, i.e., we compare $\frac{c_2}{\theta_1}$, $\frac{c_1}{\theta_2}$, and $\frac{c_1}{\theta_3}$. This could lead to resulting in investing in another feature, for example, if we have $\frac{c_1}{\theta_2} < \frac{c_2}{\theta_1} < \frac{c_1}{\theta_3}$, we would choose the second feature and invest B_2 in that feature. We continue this reasoning until we have reached the boundary. We designed our user study so the participants did not need to calculate if they reached the decision boundary and had all participants spend all their budgets.

As seen in Figure 4, the quadratic cost movement differs from norm-2 movement, which moves the point to the closest point on the decision boundary. The piece-wise function we use for our user study is similar to a quadratic cost function with $c_2 = 0.85c_1$ and a decision boundary $0.78x_1 + 0.22x_2 = 70$ for the two-dimensional case.