Beyond Rational Illusion: Behaviorally Realistic Strategic Classification

Anonymous Authors[1]

Abstract

Strategic classification (SC) studies the interaction between decision models and agents who strategically manipulate their features for favorable outcomes. Existing SC frameworks typically rely on the idealized assumption that agents are strictly rational. However, evidence from behavioral economics and psychology consistently shows that real-world decision-making is often shaped by cognitive biases, deviating from pure rationality. To formalize this limitation, we identify and define a new problem setting, termed the *behaviorally realistic strategic classification problem*, where agents' strategic manipulations deviate from full rationality due to psychological biases. Motivated by the identified limitation, we propose the **Prospect-Guided Strategic Framework** (Pro-SF) to address the problem, a principled framework grounded in prospect theory to model and learn under behaviorally realistic strategic responses. Specifically, to capture behaviorally realistic strategic manipulations, our framework reformulates the Stackelberg-style interaction between agents and the decision-maker by incorporating three key mechanisms inspired by prospect theory, including the asymmetry between benefits and costs, different subjective reference points, and non-rational probability distortion. Experiments on synthetic and real-world datasets establish Pro-SF as a behaviorally grounded approach to strategic classification, bridging machine learning and behavioral economics for more reliable deployment in the real world.

1. Introduction

Machine learning models are playing an increasingly critical role in diverse human-serving domains, such as hiring (Sánchez-Monedero et al., 2020), credit scoring (Jagtiani & Lemieux, 2019), and college admissions (Kuvcak et al., 2018). In such settings, individuals may strategically modify their observable features to obtain favorable decisions. This phenomenon reflects Goodhart's Law (Strathern, 1997): "*When a measure becomes a target, it ceases to be a good measure*." Once a model's decision rule becomes known or anticipated, agents often engage in 'gaming' behaviors to manipulate outcomes. For example, a loan applicant might temporarily inflate their reported income to appear more creditworthy. To anticipate such manipulations, strategic classification (SC) (Hardt et al., 2016) provides a specialized machine learning (ML) framework by considering the Stackelberg-style interaction between decision makers and strategic agents (Ghalme et al., 2021; Singh & Kulkarni, 2024; Chen et al., 2020), serving as a key bridge between ML and social sciences.

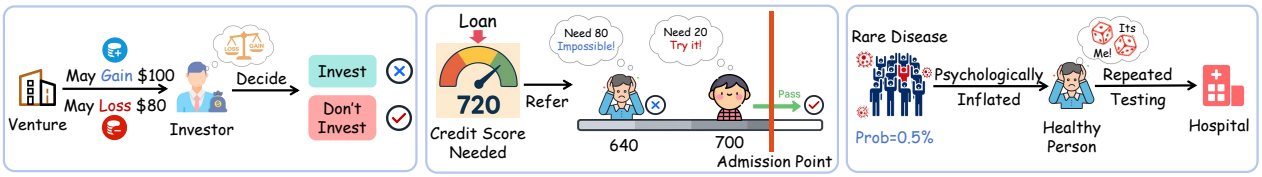
Given its growing influence, it is increasingly essential to critically examine the foundational principles of SC. To be specific, the framework of SC rests on a simplified assumption that *agents are perfectly rational* (Hardt et al., 2016; Milli et al., 2019). Consequently, under this view, an agent modifies features if, and only if, the expected benefit outweighs the cost. However, like a double-edged sword, the assumption of fully rational agents might result in conflicts with realistic scenarios, as individuals often behave in ways that deviate sharply from rational utilities:

- **Example 1.** In financial investment (Banerji et al., 2020), individuals often react more strongly to a potential \$80 loss than to a potential \$100 gain of equal probability. The same magnitude of outcome thus triggers disproportionate behavioral responses.
- **Example 2.** In credit scoring (Banerji et al., 2020), consider loan approval requires applicants to exceed a threshold A. Those whose subjective reference point B is just below A tend to make marginal adjustments, whereas those far below the threshold usually forgo effort.
- **Example 3.** In disease screening (Dwyer et al., 2022), consider a rare disease with a true prevalence of only 0.5%. Although the objective probability of infection is negligible, some individuals persist in seeking repeated testing, subjectively inflating the small chance of illness.

Unfortunately, the non-rational behaviors characterized in

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.



(a) Financial Investment (Loss Aversion) (b) Credit Scoring (Reference Bias) (c) Disease Screening (Probability Distortion)

Figure 1. Illustrative real-life scenarios of behavioral biases: (a) financial investment shaped by loss aversion, (b) credit scoring influenced by reference bias, and (c) disease screening affected by probability distortion.

the three examples, i.e., asymmetry between the benefit and cost, behaviors based on different reference points, and distorted probability towards events, are not considered in conventional SC frameworks. More broadly, extensive evidence from behavioral economics and psychology shows that real-world decision making systematically deviates from rational optimization (Carroll & Johnson, 1990; Tversky & Kahneman, 1992; Jones, 1999; Ariely & Jones, 2008). Therefore, we highlight that the rational-agent assumption, while mathematically convenient, often collapses in real-world settings. This mismatch motivates the need for a refined SC framework to capture such realistic behavioral deviations more precisely, which raises a central research challenge:

> *Can existing SC methods account for strategic manipulations shaped by psychological biases, beyond the conventional rational-utility paradigm?*

In response to this question, we first define **behaviorally realistic strategic classification** *(BR-SC)* as a new problem where agents' manipulations reflect realistic behavioral patterns driven by psychological biases, rather than fully rational best-response rules. Accordingly, the classifier is designed to account for these manipulations, ensuring robustness in deployment. Fortunately, prospect theory (Kahneman & Tversky, 2013) and followed behavioral economics highlight three pervasive mechanisms for realistic behaviors: (1) ***Loss aversion***. Individuals tend to weigh potential losses more heavily than equivalent gains (see Fig. 1(a)). (2) ***Reference bias***. Individual decision depends on subjective reference points, which differ across individuals based on their personal circumstances, expectations, or prior experiences (Fig. 1(b)). (3) ***Probability distortion***. Individuals overweight small-probability events, showing a tendency to "gamble" on unlikely opportunities (see Fig. 1(c)).

These behavioral mechanisms result in two consequential failure modes for existing SC methods, *over-defense* and *under-defense*, which degrade robustness of decision models for SC (see detailed derivation in Sec. 4). Therefore, we propose the **Prospect-Guided Strategic Framework** *(Pro-SF)* to address the BR-SC problem. Pro-SF introduces a paradigm shift for SC by integrating three key principles (i.e., asymmetry between gains and losses, subjective reference points, and probability distortion) into the Stackelberg game framework of SC. This paradigm shift captures how agents actually manipulate their features in practice and redefines both the dynamics of agents' strategic manipulation and the decision maker's optimization objective.

Our main contributions are summarized as follows:

- We identify and formalize the limitations of the rational-agent assumption in strategic classification by introducing a new problem setting, termed *behaviorally realistic strategic classification* (BR-SC). We further characterize two failure modes induced by rational-agent modeling, namely *over-defense* and *under-defense*.
- We propose the **prospect-guided strategic framework** *(Pro-SF)*, which integrates loss aversion, reference bias, and probability distortion into the Stackelberg game framework, providing a more realistic solution for strategic classification.
- We conduct extensive experiments on synthetic and real-world datasets to demonstrate that Pro-SF achieves robust performance across diverse behavioral regimes. Ablation studies and sensitivity analyses further illustrate how each behavioral component contributes to robustness.

2. Related Work

2.1. Strategic Classification

Strategic classification (SC) studies how individuals manipulate features to obtain favorable decisions (Hardt et al., 2016). Prior work largely focuses on designing decision rules that are robust to such manipulations (Dong et al., 2017; Shavit et al., 2020; Chen et al., 2020; Harris et al., 2021; Zrnic et al., 2021; Tsirtsis et al., 2024), as well as leveraging strategic behavior to incentivize genuine improvement via causal or action-based formulations (Miller et al., 2020; Chen et al., 2023; Horowitz & Rosenfeld, 2023; Vo et al., 2024; Efthymiou et al., 2025; Chang et al., 2024). Related settings include performative prediction, where repeated deployment alters the data distribution (Perdomo et al., 2020; Rosenfeld et al., 2020; Hardt et al., 2022; Mendler-Dünner et al., 2022), and societal-level regulation of strategic behavior, e.g., optimizing long-term welfare or mitigating demographic disparities (Haghtalab et al., 2020; Estornell et al., 2023a; Xie & Zhang, 2024; Zhang et al., 2022; Estornell et al., 2023b; Keswani & Celis, 2023). Re-

cent evidence suggests that human strategic manipulation deviates substantially from the rational assumption (Ebrahimi et al., 2025; Xie et al., 2025), motivating our behaviorally realistic formulation (BR-SC) and the prospect-theoretic framework (Pro-SF). **More related work** on strategic classification is included in Appendix A.1.

2.2. Behavioral Economics in Machine Learning

Behavioral economics (Mullainathan & Thaler, 2000) challenges the classical assumption that agents act as perfectly rational utility maximizers, emphasizing instead that choices are systematically shaped by cognitive biases and contextual factors (Tversky & Kahneman, 1992; Ariely & Jones, 2008). A central framework crystallizing these insights is *prospect theory* (Tversky & Kahneman, 1992; Kahneman & Tversky, 2013), which models how individuals evaluate outcomes relative to reference points, exhibit asymmetric sensitivity to gains and losses, and distort probabilities. These principles have influenced multiple domains: in economics and social science, they inform finance, consumer behavior, and public policy (Borkar & Chandak, 2021; Mercer, 2005; Vis, 2011); in computational settings, they shape reinforcement learning (Shen et al., 2014; Jie et al., 2015), mechanism design (Kuvcak et al., 2018; Leoneti & Gomes, 2021), and resource allocation (Holmes Jr et al., 2011). **More related work** is included in Appendix A.3.

3. Preliminary

We briefly review the strategic classification (SC) paradigm. Random variables are denoted by uppercase letters (e.g., X, Y), their realizations by lowercase (e.g., x, y), and boldface for vectors or matrices (e.g., $\mathbf{x}, \mathbf{X}$).

3.1. Rational Strategic Classification Model

The strategic classification problem is modeled as a Stackelberg game (Li & Sethi, 2017), where a **decision maker** defines a classification function $f : \mathbb{R}^d \to \{-1, 1\}$, and **decision subjects** (agents) strategically manipulate their features from $\mathbf{x}$ to $\mathbf{x}'$ at a cost $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ (Hardt et al., 2016; Miller et al., 2020). The optimal manipulated feature $\mathbf{x}'$ is determined by the best-response function $b_R(\mathbf{x})$:

Definition 3.1 (Rational Strategic Manipulation)**.** *The optimal modified feature vector $\mathbf{x}'$ is determined by:*

$$\mathbf{x}' = b_R(\mathbf{x}) = \arg\max_{\mathbf{x}' \in \mathcal{X}} \left[f(x') - \lambda c(x, x') \right], \tag{1}$$

where $f(x') \in \{-1, 1\}$ is the classification result after modification, $c(x, x')$ is the manipulation cost, $\lambda > 0$ is a trade-off parameter, and $\mathcal{X}$ is the feature space. Usually, the cost of manipulation is modeled as the Mahalanobis distance (Gavish et al., 2021; Chen et al., 2023).

From the decision maker's perspective, the classification rule f is designed to remain robust under such strategic manipulation:

Definition 3.2 (Decision Optimization)**.** *To mitigate manipulation, the decision maker optimizes f to maximize expected accuracy against strategic manipulation:*

$$f^* \in \arg\max_{f \in \mathcal{F}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\mathbb{1}\big(f(b_R(\mathbf{x})) = y\big) \right], \tag{2}$$

where $\mathcal{F}$ is the set of all feasible classification rules, $\mathbb{1}$ denotes the indicator function, and y is the observed label.

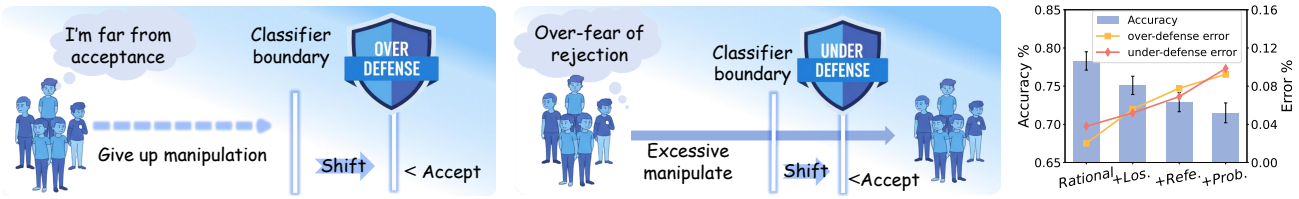
3.2. The Achilles' Heel of Strategic Classification

The rational-agent assumption underlying Eq. (1) and Eq. (2) is mathematically convenient but behaviorally restrictive. Concretely, a large body of work in behavioral economics and psychology shows that human decision making systematically departs from perfect rationality (Tversky & Kahneman, 1992; Ariely & Jones, 2008; Kahneman & Tversky, 2013; Borkar & Chandak, 2021). Among these, **Prospect Theory** (Tversky & Kahneman, 1992; Kahneman & Tversky, 2013) provides a unifying framework, capturing how individuals evaluate uncertain outcomes. Accordingly, we focus on three particularly relevant deviations from rationality that serve as the foundation for our behaviorally realistic formulation:

- **Loss aversion in manipulation.** Agents subjectively inflate perceived effort (e.g., manipulation cost) relative to gains. Consequently, agents may give up manipulations that would be beneficial under a rational-utility model because the perceived loss outweighs the perceived benefit.
- **Reference bias.** Agents evaluate outcomes relative to a subjective reference point (e.g., an expected outcome) rather than an absolute term. This means the utility of a successful manipulation varies across individuals, violating the rational-utility assumption commonly used in SC.
- **Probability distortion.** Agents may distort perceived acceptance probabilities, e.g., overweighting small probabilities. As a result, they can misjudge how much manipulation is needed to cross the decision threshold, leading to sub-optimal or insufficient adjustments.

4. Theory: Behavioral Mismatch and Failure Modes

We now provide a theoretical analysis of why classifiers trained under the rational-agent assumption can systematically fail in practice due to behavioral mismatch. We then formalize two characteristic failure modes: *over-defense* and *under-defense*.



(a) Over-defense *(b)* Under-defense *(c)* Effect of Deviations

Figure 2. Illustration of two failure modes induced by the rational-agent assumption. (a) Over-defense caused by agents giving up manipulation. (b) Under-defense caused by excessive manipulation. (c) Effects of cumulative behavioral deviations on a rational-based classifier (*Los.* = loss aversion, *Refe.* = reference bias, *Prob.* = probability distortion).

4.1. Behavioral Mismatch and Deployment Bias

Without loss of generality, the agent's strategic manipulation can be summarized by a behavioral mapping T_θ induced by a utility $U_\theta(x, x')$:

$$T_\theta : \; x \mapsto x' \in \arg\max_{x'} U_\theta(x, x'), \tag{3}$$

which induces a post-manipulation distribution D_θ different from the original feature distribution.

Classical strategic classification assumes that agents respond by maximizing a rational utility, yielding a rational mapping T_θ^{rat}. This creates a **behavioral mismatch**: the classifier is trained on the post-manipulation distribution D_θ^{rat} induced by T_θ^{rat}, but is deployed under the distribution D_θ induced by T_θ. Therefore, we have the following proposition, with a detailed proof in Appendix B.

Proposition 4.1 (Irreducible deployment bias under behavioral mismatch)**.** *Let f_θ^{rat} denote a classifier optimized on D_θ^{rat}, and define the deployment error:*

$$\delta(\theta) := \mathbb{E}_{(x,y)\sim D_\theta}\left[\mathcal{L}(f_\theta^{\text{rat}}(x), y)\right] \geq 0, \tag{4}$$

where $\mathcal{L}$ is loss function. When $D_\theta \neq D_\theta^{\text{rat}}$, then optimizing under the rational-agent assumption implies $\delta(\theta) > 0$, i.e., a non-vanishing deployment error persists.

Next, we show that behavioral mismatch manifests through two *directional* and practically failure modes: *over-defense* and *under-defense*.

4.2. Over-defense

Over-defense arises when a classifier trained under the rational-agent assumption anticipates manipulations that do not occur in practice (Fig. 2a). Agents who perceive themselves as far from acceptance due to loss aversion or reference bias, may choose not to manipulate, even when manipulation would be optimal under a rational utility.

Definition 4.2 (Over-defense)**.** *Let f_R denote a classifier trained under the rational-agent assumption. f_R exhibits* over-defense *if it adjusts its decision rule to defend against strategic manipulations that are unlikely to occur at deployment, leading to degraded classification performance.*

Example 1 (Over-defense in a linear classifier)**.** Consider a linear classifier $f(x) = \text{sign}(w^\top x - \tau)$. Under the rational-agent assumption, it is trained to defend against agents who minimally manipulate their features to cross the decision boundary. If, in practice, some agents do not manipulate at all, this anticipation causes the learned threshold to shift from τ to $\tau' > \tau$. Consequently, some instances that satisfy $w^\top x \geq \tau$—and would be accepted without any manipulation—are now rejected:

$$w^\top x \geq \tau \quad \text{but} \quad w^\top x < \tau' \;\Rightarrow\; f'_R(x) = -1. \tag{5}$$

This illustrates how defending against nonexistent manipulations induces false negatives.

Over-defense shifts the decision boundary to counter manipulations that do not occur in practice, thereby converting genuinely positive instances into false negatives. We formalize the resulting accuracy degradation in the following proposition (see proof in Appendix C).

Proposition 4.3 (Accuracy Degradation from Over-defense)**.** *Under over-defense, a classifier trained with the rational-agent assumption achieves lower accuracy, i.e.,*

$$\mathbb{E}[\mathcal{L}(f_R(x), y)] \;>\; \mathbb{E}[\mathcal{L}(f_R(b_R(x)), y)], \tag{6}$$

where $\mathcal{L}$ is the loss function for classification.

4.3. Under-defense

Under-defense occurs when agents manipulate more aggressively than anticipated by a classifier trained under the rational-agent assumption (Fig. 2b). Behaviorally, probability distortion may cause agents to overestimate their chances of acceptance, while loss aversion amplifies the fear of rejection, leading them to overshoot the manipulation level predicted by rational utility maximization.

Definition 4.4 (Under-defense)**.** *A classifier f_R with the rational-agent assumption exhibits* under-defense *if it defends only against rational manipulations, while agents in practice make larger adjustments that exceed this defense, resulting in degraded classification performance.*

Example 2 (Under-defense in a linear classifier)**.** Consider a linear classifier $f(x) = \text{sign}(w^\top x - \tau)$. Under rational

assumptions, agents are expected to manipulate minimally to the decision boundary. If, in practice, some agents overshoot this endpoint, a classifier trained only against rational manipulations fails to correctly classify these instances:

$$f_R(b_R(x)) = y \quad \text{but} \quad f_R(x^*) \neq y. \tag{7}$$

Under-defense leaves parts of the manipulated feature space unprotected, allowing strategic agents to bypass the intended defense. We formalize the resulting performance degradation in the following proposition (see proof in Appendix D).

Proposition 4.5 (Accuracy degradation from under-defense)**.** *Under under-defense, a classifier trained under the rational-agent assumption fails to guard against actual strategic manipulations, leading to reduced accuracy:*

$$\mathbb{E}\big[\mathcal{L}(f_R(x^*), y)\big] \;>\; \mathbb{E}\big[\mathcal{L}(f_R(b_R(x)), y)\big], \tag{8}$$

where $b_R(x)$ denotes the rational manipulation endpoint and x^* the actual manipulated feature.

As shown in Fig. 2c, as agents are increasingly influenced by psychological biases, the performance of a classifier trained under the rational-agent assumption deteriorates.

5. Prospect-Guided Strategic Framework

5.1. Problem Formulation

To better reflect real-world strategic behavior, we formalize a practical strategic classification setting, termed **behaviorally realistic strategic classification problem** *(BR-SC)*. Unlike classical strategic classification, which assumes agents follow rational best-response manipulations, BR-SC allows the manipulations of agents to be influenced by psychological biases.

Problem 5.1 (Behaviorally Realistic Strategic Classification (BR-SC))**.** *The BR-SC problem consists of two coupled components:*

- ***Behavioral manipulation model.*** *Given a data distribution D over (x, y), a cost function c, and a classifier f, agents are modeled by a behaviorally realistic manipulation function $b_B(x)$.*
- ***Classifier learning under behavioral responses.*** *Based on the behavioral manipulation model $b_B(x)$, the decision maker aims to train a classifier $f^* \in \mathcal{F}$ that performs robustly under behaviorally biased strategic manipulations.*

Example 3 (loan approval)**.** Consider a loan approval system, in BR-SC problem, applicants may still act strategically, but their decisions are often influenced by psychological biases, such as loss aversion, subjective reference points, and distorted beliefs about approval probability. As a result, strategic responses in real-world deployment systematically deviate from idealized rational best responses, giving rise to a more realistic strategic classification problem.

To address the *BR-SC* problem, we introduce the **Prospect-Guided Strategic Framework** *(Pro-SF)*, which leverages prospect theory to capture realistic agent manipulations and reframe the learning objective for the classifier.

5.2. Modeling Prospect-Guided Utility for Agents

We specify an explicit utility-based model for agents' strategic behavior. In particular, we adopt a prospect-guided utility that incorporates three irrational mechanisms: loss aversion, probability distortion, and reference bias.

Loss aversion in manipulation. To capture loss aversion in agents' manipulations, we adopt a prospect-style value function that evaluates perceived gains and losses on asymmetric scales. Specifically, given a candidate manipulation x, the agent's subjective value is defined as

$$v(x) \;=\; v_{\text{gain}}(x)^{\alpha} \;-\; \kappa\,(v_{\text{loss}}(x))^{\beta}, \tag{9}$$

where $v_{\text{gain}}(x)$ and $v_{\text{loss}}(x)$ denote the positive and negative components of the perceived outcome, respectively. The parameters $\alpha, \beta \in (0, 1]$ encode diminishing sensitivity, while $\kappa > 1$ amplifies the perceived weight of losses.

This value function introduces an explicit asymmetry between gains and losses.

For example, when $\kappa = 1.25$, a perceived gain of 9 and a perceived loss of 8 are evaluated asymmetrically as 9 versus 10 (i.e., $\kappa \cdot 8 = 10$), causing the loss component to dominate the decision.

Probability distortion. Agents often distort probabilities rather than evaluating them objectively, overweighting small chances of success and underweighting near certainties.

Let $p \in [0, 1]$ denote the objective probability of acceptance. The agent's subjective probability $w(p)$ is modeled via the probability weighting function:

$$w(p) \;=\; \frac{p^{\gamma}}{\big(p^{\gamma} + (1-p)^{\gamma}\big)^{1/\gamma}}, \quad \gamma \in (0, 1], \tag{10}$$

where γ controls the curvature: smaller values yield a more pronounced inverse-S shape, leading to stronger overweighting of small probabilities and underweighting of large ones.

For example, with $\gamma = 0.6$, a small objective probability $p = 0.05$ is overweighted (i.e., $w(p) = 0.2 > p$), while a large objective probability $p = 0.8$ is underweighted (i.e., $w(p) = 0.6 < p$).

Remark 1. Eq. (10) follows standard probability-weighting functions used in prospect theory, which produce an inverse-S shape probability weighting (Kahneman & Tversky, 2013; Barberis et al., 2016; Gonzalez & Wu, 1999).

Reference bias. Without loss of generality, let $s(x) \in [0, 1]$ denote the system-defined likelihood of acceptance, where

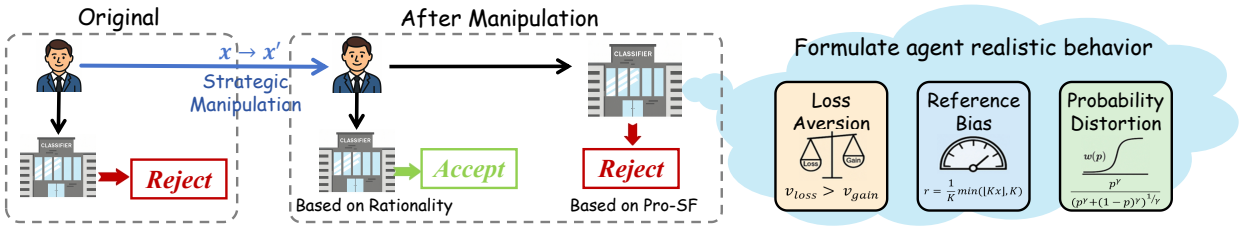


Figure 3. From rational to behaviorally realistic modeling: Pro-SF reformulates agent realistic behavior and provides robust outcomes.

$s(x) = 1$ corresponds to meeting the passing threshold. Rather than evaluating $s(x)$ precisely, agents often form a coarse-grained subjective reference point r that reflects limited sensitivity in self-assessment. We model this reference point via a quantization operator:

$$r = \frac{1}{K}\Big\lfloor K\, s(x) \Big\rfloor, \tag{11}$$

where K sets the step size $1/K$ (yielding $K+1$ discrete levels), mapping r onto equally spaced discrete levels (e.g., $K = 5$ gives $r \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$).

For example, a student after an exam may only estimate their score within a rough range (e.g., "around 70–80") rather than a precise value (e.g., 73).

Unified Prospect-Based Utility. We now unify the three behavioral components into a single prospect-based utility that governs agents' manipulation choices.

Definition 5.2 (Prospect-based utility in strategic manipulation)**.** *For an agent with subjective reference point $r \in [0, 1]$, the prospect-based utility U_{P} for manipulating features from x to x' is*

$$\begin{aligned} U_{\mathrm{P}}(x, x') =& w^{+}\big(p(x')\big)\,(1-r)^{\alpha} \\ &- \kappa\, w^{-}\big(1-p(x')\big)\, r^{\beta} - \kappa \cdot (\lambda c(x, x')^{\beta}), \end{aligned} \tag{12}$$

where $p(x') \in [0, 1]$ denotes the model-implied *probability of being classified as positive after manipulation.*

The parameters $\alpha, \beta \in (0, 1]$ capture diminishing sensitivity, $\kappa > 1$ encodes loss aversion, $\lambda > 0$ controls the strength of manipulation cost, and w^{+}, w^{-} are probability-weighting functions that distort objective probabilities into subjective decision weights.

Remark 2. We treat the manipulation cost $c(x, x')$ as an effort expenditure that contributes to the agent's subjective loss, consistent with the loss-aversion principle in prospect theory (Passarelli & Del Ponte, 2020). In special cases where effort is treated as an objective friction rather than a psychologically salient loss, the cost term can be simplified to $-\lambda\, c(x, x')^{\beta}$.

As illustrated in Fig. 3, Pro-SF reformulates agents' strategic manipulation by jointly accounting for loss aversion, reference dependence, and probability distortion.

Algorithm 1 Prospect-Guided Strategic Framework

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$; a classifier $f \in \mathcal{F}$; manipulation cost $c(\cdot)$
Parameters: $\alpha, \beta \in (0, 1]$, $\kappa > 1$, $\gamma \in (0, 1]$, $K \in \mathbb{N}$
Initialize classifier $f \in \mathcal{F}$
Perceive agents' manipulation with prospect-guided utility (Eq. (12)):
  Anticipate *loss aversion* with Eq. (9)
  Anticipate *probability distortion* with Eq. (10)
  Anticipate *reference bias* with Eq. (11)
Train classifier against manipulated data:
  Construct $\tilde{\mathcal{D}} = \{(\hat{b}_{\mathrm{P}}(x_i), y_i)\}_{i=1}^{n}$ with Eq. (13)
  Learn $f^{\star} \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}(f(\hat{b}_{\mathrm{P}}(\mathbf{x})), y)\big]$
Return $f^{\star}$

5.3. Integrating Pro-SF into Strategic Classification

We finally specify the strategic-response model and the learning objective under Pro-SF. The whole process of Pro-SF is illustrated in Algm. 1.

Definition 5.3 (Prospect-guided Strategic Manipulation)**.** *Given a classifier $f \in \mathcal{F}$ and data distribution $\mathcal{D}$, an agent strategically manipulates features in response to f:*

$$\hat{b}_{\mathrm{P}}(\mathbf{x}) \in \arg\max_{\mathbf{x}' \in \mathcal{X}} U_{\mathrm{P}}(\mathbf{x}, \mathbf{x}'), \tag{13}$$

where U_{P} is the prospect-based utility defined in Eq. (12).

Definition 5.4 (Learning objective of Pro-SF)**.** *The decision maker trains a classifier f that minimizes the expected classification loss under prospect-guided manipulation:*

$$f^{*} \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\mathcal{L}(f(\hat{b}_{\mathrm{P}}(\mathbf{x})), y)\big], \tag{14}$$

where $\mathcal{L}$ is a loss function for classification.

Learnability of behavioral parameters. The behavioral parameters $\phi = \{\alpha, \beta, \kappa, \gamma\}$ used in Pro-SF is inferred from observed manipulation behavior pairs $\{x_i, x_i'\}_{i=1}^{n}$ via maximum likelihood:

$$\phi^{*} = \arg\max_{\phi} \sum_{i} \log P_{\phi}(x_i' \mid x_i), \tag{15}$$

where $P_{\phi}(x_i' \mid x_i)$ is modeled using a discrete-choice (softmax) likelihood over a finite candidate set of feasible manipulations (see details in Appendix I).

Table 1. Performance (%) of rational-based models and Pro-SF models within different agent manipulation paradigms.

Classifier	Manipulation	Datasets					
		Adult	*Credit*	*Diabetes*	*German*	*Spam*	*Synthetic*
Rational-based	*Rational*	$78.53_{\pm 0.87}$	$76.12_{\pm 1.34}$	$70.25_{\pm 1.52}$	$74.31_{\pm 1.41}$	$83.87_{\pm 1.21}$	$81.62_{\pm 1.28}$
	Non-rational	$72.42_{\pm 1.85}$	$69.54_{\pm 2.11}$	$64.73_{\pm 1.89}$	$68.21_{\pm 2.03}$	$78.38_{\pm 1.94}$	$73.20_{\pm 2.22}$
	Mixed Behavior	$74.31_{\pm 1.51}$	$72.12_{\pm 1.87}$	$66.81_{\pm 1.65}$	$70.43_{\pm 1.92}$	$78.16_{\pm 1.63}$	$75.15_{\pm 1.74}$
Pro-SF (ours)	*Rational*	$75.31_{\pm 0.93}$	$77.02_{\pm 1.25}$	$71.23_{\pm 1.43}$	$75.18_{\pm 1.38}$	$82.92_{\pm 1.18}$	$80.41_{\pm 1.12}$
	Non-rational	$\mathbf{81.50}_{\pm 1.23}$	$\mathbf{82.41}_{\pm 1.37}$	$\mathbf{74.52}_{\pm 1.26}$	$\mathbf{79.35}_{\pm 1.42}$	$\mathbf{89.34}_{\pm 1.27}$	$\mathbf{85.20}_{\pm 1.31}$
	Mixed Behavior	$\mathbf{78.68}_{\pm 1.77}$	$\mathbf{79.13}_{\pm 1.63}$	$\mathbf{72.01}_{\pm 1.58}$	$\mathbf{76.24}_{\pm 1.71}$	$\mathbf{86.71}_{\pm 1.52}$	$\mathbf{82.34}_{\pm 1.56}$

Table 2. Performance of our ablation study on behavioral mechanisms.

Classifier	Behavioral Factors			Metrics		
	Refe.	*Prob.*	*Los.*	*Accuracy* (%) ↑	*Over-defense error* (%) ↓	*Under-defense error* (%) ↓
f_{Pro-sf}	✓	✓	✓	78.92 ± 0.13	5.17 ± 0.11	3.22 ± 0.09
$f_{Refe+Prob}$	✓	✓	✗	77.72 ± 0.27	6.24 ± 0.08	5.29 ± 0.12
$f_{Refe+Los}$	✓	✗	✓	77.80 ± 0.16	7.21 ± 0.10	5.05 ± 0.07
$f_{Prob+Los}$	✗	✓	✓	77.45 ± 0.09	6.28 ± 0.06	4.79 ± 0.13
f_{Refe}	✓	✗	✗	75.33 ± 0.11	9.12 ± 0.09	7.27 ± 0.10
f_{Prob}	✗	✓	✗	73.50 ± 0.15	9.26 ± 0.08	9.23 ± 0.12
f_{Los}	✗	✗	✓	75.91 ± 0.11	7.14 ± 0.11	7.05 ± 0.09

Note: 1) *Refe.* = Reference bias. 2) *Prob.* = Probability distortion. 3) *Los.* = Loss aversion.

6. Experiment

6.1. Experimental Setup

Dataset. We evaluate our framework on five datasets, including four real-world and one synthetic benchmarks: *Credit*, *Adult*, *Diabetes*, *German*, *Spam*, and *Synthetic* (see detailed description in Appendix G.1).

Agent behavior paradigms. To evaluate robustness under different behavioral assumptions, we consider three strategic manipulations:

- **Fully rational.** The classical strategic classification paradigm, where agents manipulate their features to maximize utility in Eq. (1) (Hardt et al., 2016).
- **Non-rational.** Agents stochastically deviate from the rational best-response strategy, driven by behavioral mechanisms in behavioral economics (Kahneman & Tversky, 2013; Kuvcak et al., 2018; Leoneti & Gomes, 2021).
- **Mixed behavioral.** A heterogeneous population in which a proportion π of agents behave fully rationally, while the remaining $(1-\pi)$ exhibit non-rational strategic behavior.

Metric. We use **accuracy** as the primary metric and introduce two additional metrics: **over-defense error** *(ODE)* and **under-defense error** (*UDE*). ODE measures false rejections caused by the classifier being overly defensive. UDE measures false acceptances due to non-rational manipulation of agents and insufficient defense of the classifier.

Baseline. We mainly conduct experiments with linear models, consistent with previous standard approaches (Chen et al., 2023; Shavit et al., 2020; Ghalme et al., 2021) with Mahalanobis distance for manipulation cost (Gavish et al., 2021) (other forms included in Appendix K).

Implementation. We implement the classifier f as a logistic regression model trained with cross-entropy loss and learning rate 10^{-3}. In prospect-theoretic utility, we set the curvature parameters $\alpha = 0.8$ and $\beta = 0.7$, the loss aversion coefficient $\kappa = 2.25$, and the probability-weighting parameter $\gamma = 0.7$ by default. In the mixed agent behavior paradigm, the proportion is set as $\pi = 0.2$. More implementation details are included in Appendix G.3

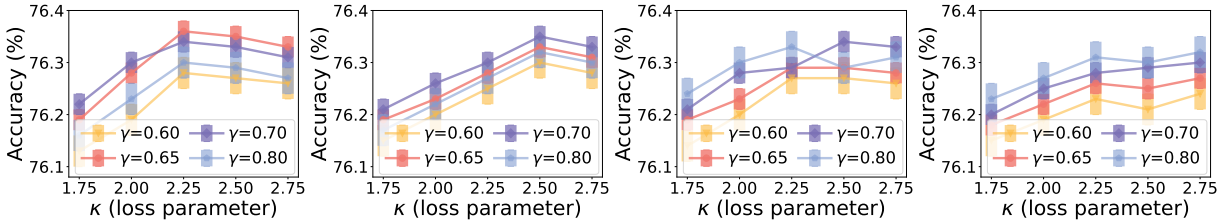
6.2. Ablation Study

To better understand the design of our Pro-SF, we conduct the following ablation studies.

Ablation on behavioral mechanisms. We first examine the necessity of the three core behavioral mechanisms: loss aversion, probability weighting, and reference bias. Starting from the full model, we construct variants by *neutralizing* one component at a time. For example, the **rational-weighting** variant removes probability distortion by setting $w^+(p) = p$ and $w^-(1-p) = 1-p$.

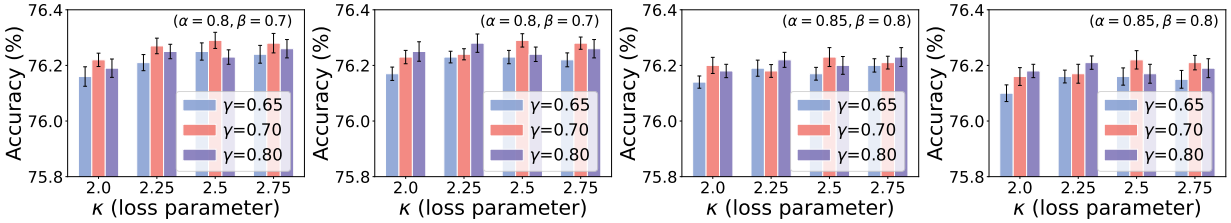
Parameter sensitivity. [1] We further assess robustness to

[1]The tested parameter ranges are determined based on our parameter learning objective (Eq. (15)) and ranges commonly adopted in prospect-theoretic and behavioral economics literature (Kahneman & Tversky, 2013; Borkar & Chandak, 2021).



(a) $\alpha = 0.8,\ \beta = 0.8$ *(b)* $\alpha = 0.85,\ \beta = 0.8$ *(c)* $\alpha = 0.85,\ \beta = 0.7$ *(d)* $\alpha = 0.8,\ \beta = 0.7$

Figure 4. Parameter ablation results with parameters $(\alpha, \beta, \kappa, \gamma)$ and $r \in \{0, 0.2, 0.4, 0.6, 0.8\}$.



(a) $r \in \{0, 0.3, 0.6, 0.9\}$ *(b)* $r \in \{0, 0.4, 0.7\}$ *(c)* $r \in \{0, 0.3, 0.6, 0.9\}$ *(d)* $r \in \{0, 0.4, 0.7\}$

Figure 5. Parameter ablation results with parameters κ, γ, r with different (α, β).

parameter choices by varying one parameter (or parameter pair). Specifically, we consider:

- $(\alpha, \beta) \in \{(0.85, 0.8), (0.85, 0.75), (0.8, 0.8), (0.8, 0.7)\}$;
- $\kappa \in \{1.75, 2.0, 2.25, 2.5, 2.75\}$;
- $\gamma \in \{0.6, 0.65, 0.7, 0.8\}$,
- r determined by the reference granularity $K \in \{3, 4, 5\}$.

A notation table is provided in Appendix G.2 (Table 4) and more results are included in Appendix G.4(Fig. 6).

An alternative of probability distortion modeling is the two-parameter *Prelec* function (Prelec, 1998), $w(p) = \exp\big(-\eta(-\ln p)^{\phi}\big)$, $\phi, \eta > 0$. We examine this function in Appendix G.4 (Table 6).

6.3. Result and Analysis

Overall performance. As shown in Table 1, Pro-SF consistently outperforms rational-based models when agents exhibit behavioral deviations, both in the non-rational and the more realistic mixed settings, across all datasets. At the same time, Pro-SF maintains competitive performance under the fully rational paradigm, indicating that incorporating behavioral modeling does not compromise performance when classical assumptions hold. Overall, these results demonstrate that Pro-SF provides robust performance across diverse strategic environments.

Ablation on behavioral mechanisms. Table 2 shows that each component contributes meaningfully, and the full Pro-SF (all three enabled) consistently performs best. *Reference bias* yields the largest marginal gain: removing it hurts most because the model can no longer capture abstention, which is key to correcting over-defense. Dropping *probability weighting* reduces robustness to distorted tail probabilities and increases under-defense, while removing *loss aversion* eliminates gain–loss asymmetry and weakens boundary stability. Overall, multi-mechanism variants outperform single-mechanism ones, indicating strong complementarity: reference bias governs *whether* agents move, whereas probability weighting and loss aversion determine *how far* they move.

Parameter sensitivity. Fig. 4 and Fig. 5 examine the effect of varying the Pro-SF parameters. Overall, accuracy remains stable across wide ranges, showing that Pro-SF does not rely on fine-tuned hyperparameters. Specifically, adjusting the loss aversion coefficient κ only changes outcomes marginally, suggesting robustness to different levels of risk sensitivity. Variations in the probability distortion parameter γ shift the relative emphasis on tail events but do not alter the overall trend. Different bins of r only cause some fluctuations in performance, but are all better than the rational classifier. Finally, different curvature settings (α, β) yield consistent results, confirming that Pro-SF maintains effectiveness under diverse utility shapes.

7. Conclusion

This work challenges the classical rational-agent assumption in strategic classification and formalizes the problem of *behaviorally realistic strategic classification*. We theoretically characterize the impact of behavioral mismatch on deployment performance, and propose the *Prospect-Guided Strategic Framework*, which incorporates three key psychological mechanisms underlying strategic manipulation and provides a behaviorally grounded solution for SC in the real-world. Future work will extend this framework to richer and more diverse deployment settings.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Ariely, D. and Jones, S. *Predictably irrational*. HarperCollins New York, 2008.

Banerji, J., Kundu, K., and Alam, P. A. An empirical investigation into the influence of behavioral biases on investment behavior. *SCMS Journal of Indian Management*, 17 (1):81–98, 2020.

Barberis, N., Mukherjee, A., and Wang, B. Prospect theory and stock returns: An empirical test. *The review of financial studies*, 29(11):3068–3107, 2016.

Barberis, N., Jin, L. J., and Wang, B. Prospect theory and stock market anomalies. *The Journal of Finance*, 76(5): 2639–2687, 2021.

Becker, B. and Kohavi, R. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Borkar, V. S. and Chandak, S. Prospect-theoretic q-learning. *Systems & Control Letters*, 156:105009, 2021.

Carroll, J. S. and Johnson, E. J. *Decision research: A field guide*. Sage Publications, Inc, 1990.

Chang, T., Warrenburg, L., Park, S.-H., Parikh, R., Makar, M., and Wiens, J. Who's gaming the system? a causally-motivated approach for detecting strategic adaptation. *Advances in Neural Information Processing Systems*, 37: 42311–42348, 2024.

Chen, Y., Liu, Y., and Podimata, C. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems*, 33:15265–15276, 2020.

Chen, Y., Wang, J., and Liu, Y. Learning to incentivize improvements from strategic agents. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=W98AEKQ38Y.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences, 2017. URL https://arxiv.org/abs/1710.07887.

Dwyer, A. A., Uveges, M. K., Dockray, S., and Smith, N. Exploring rare disease patient attitudes and beliefs regarding genetic testing: implications for person-centered care. *Journal of Personalized Medicine*, 12(3):477, 2022.

Ebrahimi, R., Vaccaro, K., and Naghizadeh, P. The double-edged sword of behavioral responses in strategic classification: Theory and user studies. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 868–886, 2025.

Efthymiou, V., Podimata, C., Sen, D., and Ziani, J. Incentivizing desirable effort profiles in strategic classification: The role of causality and uncertainty. *arXiv preprint arXiv:2502.06749*, 2025.

Eilat, I., Finkelshtein, B., Baskin, C., and Rosenfeld, N. Strategic classification with graph neural networks. *arXiv preprint arXiv:2205.15765*, 2022.

Estornell, A., Chen, Y., Das, S., Liu, Y., and Vorobeychik, Y. Incentivizing recourse through auditing in strategic classification. In *IJCAI*, 2023a.

Estornell, A., Das, S., Liu, Y., and Vorobeychik, Y. Group-fair classification with strategic agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pp. 389–399, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594006. URL https://doi.org/10.1145/3593013.3594006.

Evans, B. P. and Prokopenko, M. Bounded rationality for relaxing best response and mutual consistency: the quantal hierarchy model of decision making. *Theory and Decision*, 96(1):71–111, 2024.

Gavish, M., Talmon, R., Su, P.-C., and Wu, H.-T. Optimal recovery of precision matrix for mahalanobis distance from high dimensional noisy observations in manifold learning, 2021. URL https://arxiv.org/abs/1904.09204.

Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., and Rosenfeld, N. Strategic classification in the dark, 2021. URL https://arxiv.org/abs/2102.11592.

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

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Z. Maximizing welfare with incentive-aware evaluation mechanisms. *arXiv preprint arXiv:2011.01956*, 2020.

Hardt, M. and Mendler-Dünner, C. Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*, 2023.

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

Hardt, M., Jagadeesan, M., and Mendler-Dünner, C. Performative power. *Advances in Neural Information Processing Systems*, 35:22969–22981, 2022.

Harris, K., Heidari, H., and Wu, S. Z. Stateful strategic regression. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28728–28741. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f1404c2624fa7f2507ba04fd9dfc5fb1-Paper.pdf.

Harris, K., Ngo, D. D. T., Stapleton, L., Heidari, H., and Wu, S. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*, pp. 8502–8522. PMLR, 2022.

Hofmann, H. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Holmes Jr, R. M., Bromiley, P., Devers, C. E., Holcomb, T. R., and McGuire, J. B. Management theory applications of prospect theory: Accomplishments, challenges, and opportunities. *Journal of Management*, 37(4):1069–1107, 2011.

Hopkins, Mark, Reeber, Erik, Forman, George, Suermondt, and Jaap. Spambase. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C53G6X.

Horowitz, G. and Rosenfeld, N. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*, pp. 13233–13253. PMLR, 2023.

Hossain, S., Micha, E., Chen, Y., and Procaccia, A. Strategic classification with externalities. *arXiv preprint arXiv:2410.08032*, 2024.

Jagtiani, J. and Lemieux, C. The roles of alternative data and machine learning in fintech lending: evidence from the lendingclub consumer platform. *Financial Management*, 48(4):1009–1029, 2019.

Jie, C., Fu, M., Marcus, S., Szepesvári, C., et al. Cumulative prospect theory meets reinforcement learning: Prediction and control. *arXiv preprint arXiv:1506.02632*, 2015.

Jones, B. D. Bounded rationality. *Annual review of political science*, 2(1):297–321, 1999.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.

Keswani, V. and Celis, L. E. Addressing strategic manipulation disparities in fair classification. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–11, 2023.

Kleinberg, J. and Raghavan, M. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Kuvcak, D., Jurivcic, V., and DJambic, G. Machine learning in education-a survey of current research trends. *Annals of DAAAM & Proceedings*, 29, 2018.

Laibson, D. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.

Leoneti, A. B. and Gomes, L. F. A. M. A novel version of the todim method based on the exponential model of prospect theory: The exptodim method. *European Journal of Operational Research*, 295(3):1042–1055, 2021. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2021.03.055. URL https://www.sciencedirect.com/science/article/pii/S0377221721002812.

Levanon, S. and Rosenfeld, N. Strategic classification made practical. In *International Conference on Machine Learning*, pp. 6243–6253. PMLR, 2021.

Li, T. and Sethi, S. P. A review of dynamic stackelberg game models. *Discrete & Continuous Dynamical Systems-Series B*, 22(1), 2017.

Lopez-Rojas, E., Elmir, A., and Axelsson, S. Paysim: A financial mobile money simulator for fraud detection. In *28th European modeling and simulation symposium, EMSS, Larnaca*, pp. 249–255. Dime University of Genoa, 2016.

Mendler-Dünner, C., Ding, F., and Wang, Y. Anticipating performativity by predicting from predictions. *Advances in neural information processing systems*, 35:31171–31185, 2022.

Mercer, J. Prospect theory and political science. *Annu. Rev. Polit. Sci.*, 8(1):1–21, 2005.

Miller, J., Milli, S., and Hardt, M. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pp. 230–239, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3287560.3287576.

Mofakhami, M., Mitliagkas, I., and Gidel, G. Performative prediction with neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 11079–11093. PMLR, 2023.

Mullainathan, S. and Thaler, R. H. Behavioral economics. Working Paper w7948, National Bureau of Economic Research, October 2000.

O'donoghue, T. and Rabin, M. Doing it now or later. *American economic review*, 89(1):103–124, 1999.

Passarelli, F. and Del Ponte, A. Prospect theory, loss aversion, and political behavior. In *Oxford Research Encyclopedia of Politics*. 2020.

Payne, K. An analysis of ai decision under risk: Prospect theory emerges in large language models. *arXiv preprint arXiv:2508.00902*, 2025.

Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

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

Rosenfeld, N., Hilgard, A., Ravindranath, S. S., and Parkes, D. C. From predictions to decisions: Using lookahead regularization. *Advances in Neural Information Processing Systems*, 33:4115–4126, 2020.

Sánchez-Monedero, J., Dencik, L., and Edwards, L. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 458–468, 2020.

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

Shen, Y., Tobia, M. J., Sommer, T., and Obermayer, K. Risk-sensitive reinforcement learning. *Neural computation*, 26 (7):1298–1328, 2014.

Singh, M. K. and Kulkarni, A. A. Optimal stochastic decision rule for strategic classification. In *2024 National Conference on Communications (NCC)*, pp. 1–6. IEEE, 2024.

Strathern, M. 'improving ratings': audit in the british university system. *European Review*, 5(3):305–321, 1997. doi: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

Teboul, A. Diabetes health indicators dataset, 2015. URL https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

Todd, P. M. and Gigerenzer, G. Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, 23 (5):727–741, 2000.

Tsirtsis, S., Tabibian, B., Khajehnejad, M., Singla, A., Schölkopf, B., and Gomez-Rodriguez, M. Optimal decision making under strategic behavior. *Management Science*, 2024.

Tversky, A. and Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

Vis, B. Prospect theory and political decision making. *Political Studies Review*, 9(3):334–343, 2011.

Vo, K. Q., Aadil, M., Chau, S. L., and Muandet, K. Causal strategic learning with competitive selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15411–15419, 2024.

Xie, T. and Zhang, X. Non-linear welfare-aware strategic learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 1660–1671, 2024.

Xie, T., Rauch, P., and Zhang, X. How strategic agents respond: Comparing analytical models with llm-generated responses in strategic classification. *arXiv preprint arXiv:2501.16355*, 2025.

Yeh, I.-C. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

Zhang, X., Khalili, M. M., Jin, K., Naghizadeh, P., and Liu, M. Fairness interventions as (Dis)Incentives for strategic manipulation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26239–26264. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhang22l.html.

Zrnic, T., Mazumdar, E., Sastry, S., and Jordan, M. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems*, 34: 15257–15269, 2021.

Appendix Contents

A. Additional Related Work

A.1. More Strategic Machine Learning Work

Several additional directions in strategic machine learning complement the studies highlighted in Section 2.

Robustness extensions. Beyond classical defenses, researchers have explored stochastic classifiers (Singh & Kulkarni, 2024), differentiable optimization layers for end-to-end robustness (Levanon & Rosenfeld, 2021), and graph-based models to capture inter-agent dependencies (Eilat et al., 2022). Multi-agent formulations further investigate externalities and collective dynamics in strategic settings (Hossain et al., 2024).

Positive manipulation. A growing body of work highlights the constructive potential of strategic behavior. For example, classifiers that incentivize authentic qualification gains have been proposed in education and hiring contexts (Kleinberg & Raghavan, 2020; Harris et al., 2022). These approaches complement causal frameworks (Miller et al., 2020; Chen et al., 2023), which distinguish between manipulable and improvable features.

Performative prediction extensions. Beyond the foundational contributions (Perdomo et al., 2020; Rosenfeld et al., 2020; Hardt et al., 2022; Mendler-Dünner et al., 2022), further work explores neural methods for dynamic feedback loops (Mofakhami et al., 2023) and more recent advances in model-induced distribution shifts (Hardt & Mendler-Dünner, 2023).

Fairness perspectives. In addition to direct fairness interventions (Zhang et al., 2022; Estornell et al., 2023b; Keswani & Celis, 2023), several studies consider fairness as an incentive-alignment mechanism that influences agents' strategic behavior, thereby connecting individual manipulation with broader social equity.

A.2. Position of Our Work

Our work, Pro-SF, is designed to *complement* the classical strategic classification paradigm rather than overturn it. The rational-agent model remains a useful idealization—and a natural baseline—for analyzing strategic behavior under well-specified utilities and costs. Our contribution is to provide a behaviorally grounded formulation for settings where strategic responses systematically deviate from this idealization due to psychological biases. Importantly, Pro-SF is fully compatible with classical SC: when behavioral mechanisms are absent (e.g., $\kappa = 1$ and probability weighting reduces to the identity), Pro-SF reduces to the standard rational-response formulation.

A.3. Additional Behavioral Economics Background

Beyond the works highlighted in Section 2, a broader body of literature illustrates how behavioral economics explains systematic deviations from rationality and informs computational models.

Bounded rationality and heuristics. Human decision-making is constrained by limited cognitive resources and often relies on simplified rules rather than exhaustive optimization (Todd & Gigerenzer, 2000; Evans & Prokopenko, 2024). This perspective suggests that agents may respond in partial, myopic, or context-dependent ways instead of precise best responses.

Temporal and framing effects. Individuals overweight immediate costs relative to delayed benefits and respond differently to equivalent options depending on presentation (Laibson, 1997; O'donoghue & Rabin, 1999). These effects imply that manipulations which appear objectively beneficial may still be avoided or inconsistently adopted.

Extended applications. Behavioral frameworks have been applied across diverse algorithmic domains. Examples include portfolio optimization and asset pricing in finance (Barberis et al., 2021), multi-criteria decision methods (Leoneti & Gomes, 2021), fairness-aware allocation in management science (Holmes Jr et al., 2011), and human–AI decision support systems (Payne, 2025). Collectively, these studies demonstrate the versatility of prospect-theoretic and behavioral models in explaining and predicting non-rational behavior.

B. Proof of Proposition 4.1

We prove Proposition 4.1 under a standard threshold classifier and the 0–1 loss. Recall that the behavioral response mapping T_θ induces the deployment post-manipulation distribution D_θ, while the rational response mapping T_θ^{rat} induces the training post-manipulation distribution D_θ^{rat}. Let f_θ^{rat} be a classifier optimized on D_θ^{rat}, and define the deployment error

$$\delta(\theta) := \mathbb{E}_{(x,y)\sim D_\theta}\big[\mathcal{L}(f_\theta^{\mathrm{rat}}(x), y)\big]. \tag{16}$$

B.1. Classifier Form and Loss

Let f_θ^{rat} be a threshold classifier of the form

$$f_\theta^{\mathrm{rat}}(x) = \mathbf{1}\{s_\theta(x) \geq \tau\}, \tag{17}$$

where $s_\theta : \mathcal{X} \to \mathbb{R}$ is the decision score and $\tau \in \mathbb{R}$ is the threshold. We take the 0–1 loss:

$$\mathcal{L}(f(x), y) = \mathbf{1}\{f(x) \neq y\}. \tag{18}$$

Hence,

$$\delta(\theta) = \mathbb{P}_{(x,y)\sim D_\theta}\big(f_\theta^{\mathrm{rat}}(x) \neq y\big). \tag{19}$$

B.2. Decision-relevant Mismatch Assumption

Since $D_\theta \neq D_\theta^{\mathrm{rat}}$, there exists a measurable set on which the two distributions assign different probability mass. We assume this discrepancy occurs in a decision-relevant region near the boundary and leads to unavoidable label disagreement: there exist constants $\epsilon > 0$ and $\gamma > 0$ such that

$$\mathbb{P}_{(x,y)\sim D_\theta}\Big(|s_\theta(x) - \tau| \leq \gamma,\ y \neq \mathbf{1}\{s_\theta(x) \geq \tau\}\Big) \ \geq\ \epsilon. \tag{20}$$

Intuitively, Eq. (20) states that under the true behavioral response, a non-zero fraction of deployment samples lies within a γ-neighborhood of the decision boundary but carries labels that disagree with the classifier's prediction. This is exactly the type of boundary-sensitive discrepancy induced when the training-time response model T_θ^{rat} mis-specifies the true response T_θ, producing a different post-manipulation distribution.

B.3. Lower Bound on Deployment Error

Let

$$A := \Big\{(x,y) : |s_\theta(x) - \tau| \le \gamma,\ y \neq \mathbf{1}\{s_\theta(x) \ge \tau\}\Big\}. \tag{21}$$

By definition of the 0–1 loss, on the set A we have $\mathcal{L}(f_\theta^{\mathrm{rat}}(x), y) = 1$. Therefore,

$$\begin{aligned}\delta(\theta) &= \mathbb{E}_{(x,y)\sim D_\theta}\big[\mathbf{1}\{f_\theta^{\mathrm{rat}}(x) \neq y\}\big] \\ &\ge \mathbb{E}_{(x,y)\sim D_\theta}\big[\mathbf{1}\{(x,y) \in A\}\big] \\ &= \mathbb{P}_{(x,y)\sim D_\theta}\big((x,y) \in A\big) \\ &\ge \epsilon\end{aligned}$$

Hence $\delta(\theta) \ge \epsilon > 0$, proving that the deployment-time error is strictly positive and cannot be eliminated by optimizing under the rational-agent assumption.

Under behavioral mismatch $D_\theta \neq D_\theta^{\mathrm{rat}}$ and the decision-relevant boundary-mass condition (20), we have $\delta(\theta) > 0$, i.e., a non-vanishing deployment error persists.

C. Proof of Proposition 4.3

Proof. Recall that $b_R(x)$ denotes the rational best response (Eq. (1)), and f_R^* is the classifier optimized under the rational-agent assumption, i.e., an empirical risk minimizer of

$$R_{\mathrm{train}}(f) \ = \ \mathbb{E}[\mathcal{L}(f(b_R(X)), Y)] \quad \text{(cf. Eq. (2))}. \tag{22}$$

Intuitively, f_R^* learns as if every agent will manipulate their features to the rational endpoint $b_R(x)$. However, in deployment, some agents may *not* manipulate due to behavioral biases such as reference or loss aversion. This mismatch between training and deployment is the root cause of accuracy degradation.

C.1. Defining the Abstention Set

We first isolate the problematic subset of instances:

$$A \ = \ \{x : b_R(x) \neq x \ \text{ and agents actually stay at } x\}. \tag{23}$$

That is, A contains inputs where the rational model predicts manipulation ($b_R(x) \neq x$), but in reality the agent abstains and remains at the original feature x. We assume $P(A) > 0$ so that this situation occurs with non-negligible probability.

C.2. Key Quantities on A.

On this abstention set, two factors drive the deployment–training discrepancy:

Training-side error at rational endpoints:

$$\varepsilon_R(f_R^*; A) := \Pr\big(f_R^*(b_R(X)) \neq Y \,\big|\, X \in A\big). \tag{24}$$

This is the error f_R^* makes at the rationally shifted points $b_R(X)$, which it was trained on.

Disagreement induced by boundary shift:

$$\tau_A := \Pr\big(f_R^*(X) \neq f_R^*(b_R(X)) \,\big|\, X \in A\big). \tag{25}$$

This captures how much the classifier's decision boundary has shifted to defend against $b_R(X)$. If $f_R^*(X) \neq f_R^*(b_R(X))$, then f_R^* makes a different prediction on the true input X than on the assumed manipulated input $b_R(X)$. Thus, τ_A measures the "extra disagreement region" caused by over-defense.

C.3. Conditional Decomposition.

We now compare the losses at the true point x and the rational endpoint $b_R(x)$. For any classifier f and any (x, y) with $x \in A$, we have

$$\mathcal{L}(f(x), y) - \mathcal{L}(f(b_R(x)), y) = \mathbb{1}\{f(x) \neq y,\ f(b_R(x)) = y\} - \mathbb{1}\{f(x) = y,\ f(b_R(x)) \neq y\}. \tag{26}$$

This identity simply says: the loss difference is positive if f misclassifies x but correctly classifies $b_R(x)$, and negative if the opposite happens.

Taking expectation conditional on $X \in A$, we obtain

$$\begin{aligned}&\mathbb{E}\big[\mathcal{L}(f(X), Y) - \mathcal{L}(f(b_R(X)), Y) \mid X \in A\big] \\ &\quad = \Pr(f(X) \neq Y, f(b_R(X)) = Y \mid X \in A) - \Pr(f(X) = Y, f(b_R(X)) \neq Y \mid X \in A).\end{aligned}$$

To simplify, we use two inclusions:

$$\begin{aligned}\{f(X) \neq Y,\ f(b_R(X)) = Y\} &\supseteq \{f(X) \neq f(b_R(X))\} \setminus \{f(b_R(X)) \neq Y\}, \\ \{f(X) = Y,\ f(b_R(X)) \neq Y\} &\subseteq \{f(b_R(X)) \neq Y\}.\end{aligned}$$

The first line says: whenever $f(X)$ and $f(b_R(X))$ disagree, the difference can only be negated if $b_R(X)$ is itself mislabeled. The second line says: if $f(b_R(X))$ is wrong, then one possible case is $f(X) = Y$, but this cannot exceed the entire error set.

Applying these to the conditional expectation gives a clean lower bound:

$$\begin{aligned}&\mathbb{E}\big[\mathcal{L}(f(X), Y) - \mathcal{L}(f(b_R(X)), Y) \mid X \in A\big] &(27)\\ &\qquad \geq\ \Pr(f(X) \neq f(b_R(X)) \mid X \in A) - 2\,\Pr(f(b_R(X)) \neq Y \mid X \in A). &(28)\end{aligned}$$

C.4. Apply to f_R^* and Globalize.

Now substitute $f = f_R^*$ into (27). Using the definitions of τ_A and $\varepsilon_R(f_R^*; A)$, we obtain

$$\mathbb{E}\big[\mathcal{L}(f_R^*(X), Y) - \mathcal{L}(f_R^*(b_R(X)), Y) \mid X \in A\big]\ \geq\ \tau_A - 2\,\varepsilon_R(f_R^*; A). \tag{29}$$

Multiplying both sides by $P(A)$ and adding the (unconstrained) contribution from $X \notin A$ gives the global gap:

$$\mathbb{E}[\mathcal{L}(f_R^*(X), Y)] - \mathbb{E}[\mathcal{L}(f_R^*(b_R(X)), Y)]\ \geq\ P(A)\,\big(\tau_A - 2\,\varepsilon_R(f_R^*; A)\big). \tag{30}$$

If the disagreement region dominates the residual training error on A, i.e.,

$$\tau_A > 2\,\varepsilon_R(f_R^*; A), \tag{31}$$

then the right-hand side is strictly positive. This yields the claimed inequality:

$$\mathbb{E}[\mathcal{L}(f_R^*(X), Y)]\ >\ \mathbb{E}[\mathcal{L}(f_R^*(b_R(X)), Y)]. \tag{32}$$

Thus, a classifier trained under rational-agent assumptions indeed suffers accuracy degradation in deployment due to over-defense.

Remark 3. Geometrically, f_R^* shifts its boundary outward to protect against the rational endpoints $b_R(x)$. But on the abstention set A, agents actually remain at x. This creates a disagreement region of size τ_A where f_R^*'s prediction at x differs from that at $b_R(x)$. Since f_R^* was only trained on $b_R(x)$, its prediction at x is often incorrect, and the gap $\tau_A - 2\varepsilon_R$ quantifies this excess deployment error.

D. Proof of Proposition 4.5

Proof. Let $b_R(x)$ denote the rational best response (Eq. (1)), and let f_R^* be the classifier optimized under the rational-agent assumption, i.e., an ERM of

$$R_{\text{train}}(f) = \mathbb{E}[\mathcal{L}(f(b_R(X)), Y)] \quad \text{(cf. Eq. (2))}. \tag{33}$$

Intuitively, f_R^* is trained to defend against manipulations exactly to $b_R(x)$. However, under behavioral biases such as loss aversion and probability distortion, agents may "overshoot" and move beyond $b_R(x)$. This is the essence of the *under-defense* problem: the classifier does not anticipate manipulations that go further than the rational endpoint.

D.1. Define the Overshoot Set

Let $b_B(x)$ denote the actual post-manipulation point under behavioral biases. We focus on inputs where overshooting occurs:

$$B \;=\; \Big\{x : b_R(x) \neq x \;\text{ (rationally predicted manipulation)} \tag{34}$$

$$\text{and} \quad b_B(x) \neq b_R(x) \text{ with overshoot beyond } b_R(x)\Big\}. \tag{35}$$

We assume $P(B) > 0$, so that overshoot happens with non-negligible probability. Here, "overshoot" simply means that $b_B(x)$ lies further along the manipulation direction than $b_R(x)$, which may cause different classifier outputs.

D.2. Key Quantities

On B, we define two measures:

Rational-reference error: how often the classifier errs at the rational endpoint:

$$\varepsilon_R(f; B) \;:=\; \Pr\big(f(b_R(X)) \neq Y \;\big|\; X \in B\big). \tag{36}$$

Overshoot-induced disagreement: how often predictions at $b_B(X)$ and $b_R(X)$ differ:

$$\tau_B \;:=\; \Pr\big(f_R^*(b_B(X)) \neq f_R^*(b_R(X)) \;\big|\; X \in B\big). \tag{37}$$

Intuitively, $\varepsilon_R(f_R^*; B)$ reflects how well f_R^* performs where it was trained ($b_R(x)$), while τ_B captures the risk that overshoot moves points into regions where f_R^* makes different predictions.

D.3. Conditional Decomposition.

For any classifier f and any (x, y) with $x \in B$, the loss difference between overshoot and rational endpoints is

$$\begin{aligned}&\mathcal{L}(f(b_B(x)), y) - \mathcal{L}(f(b_R(x)), y)\\ &\quad = \mathbb{1}\{f(b_B(x)) \neq y,\, f(b_R(x)) = y\} - \mathbb{1}\{f(b_B(x)) = y,\, f(b_R(x)) \neq y\}.\end{aligned}$$

Taking expectation conditional on $X \in B$ gives

$$\begin{aligned}&\mathbb{E}\big[\mathcal{L}(f(b_B(X)), Y) - \mathcal{L}(f(b_R(X)), Y) \;\big|\; X \in B\big]\\ &\quad = \Pr(f(b_B(X)) \neq Y, f(b_R(X)) = Y \mid X \in B) - \Pr(f(b_B(X)) = Y, f(b_R(X)) \neq Y \mid X \in B).\end{aligned}$$

Now, note the following inclusions:

$$\begin{aligned}&\{f(b_B) \neq Y,\, f(b_R) = Y\} \supseteq \{f(b_B) \neq f(b_R)\} \setminus \{f(b_R) \neq Y\},\\ &\{f(b_B) = Y,\, f(b_R) \neq Y\} \subseteq \{f(b_R) \neq Y\}.\end{aligned}$$

The first line says: if the predictions at b_B and b_R disagree, this difference usually contributes to loss unless b_R is already mislabeled. The second line says: if $f(b_R)$ is wrong, then the case where $f(b_B)$ is correct is bounded by that same error set.

Applying these gives the lower bound:

$$\begin{aligned}&\mathbb{E}\big[\mathcal{L}(f(b_B(X)), Y) - \mathcal{L}(f(b_R(X)), Y) \;\big|\; X \in B\big]\\ &\qquad\qquad \geq\; \Pr(f(b_B(X)) \neq f(b_R(X)) \mid X \in B) - 2\,\Pr(f(b_R(X)) \neq Y \mid X \in B).\end{aligned} \tag{38}$$

D.4. Apply to f_R^* and Globalize.

Substituting $f = f_R^*$ in (38), we obtain

$$\mathbb{E}\big[\mathcal{L}(f_R^*(b_B(X)), Y) - \mathcal{L}(f_R^*(b_R(X)), Y) \;\big|\; X \in B\big] \;\geq\; \tau_B - 2\,\varepsilon_R(f_R^*; B). \tag{39}$$

Multiplying both sides by $P(B)$ and adding the (unrestricted-sign) contribution from $X \notin B$ gives the global error gap:

$$\mathbb{E}[\mathcal{L}(f_R^*(b_B(X)), Y)] - \mathbb{E}[\mathcal{L}(f_R^*(b_R(X)), Y)] \;\geq\; P(B)\,\big(\tau_B - 2\,\varepsilon_R(f_R^*; B)\big). \tag{40}$$

If the overshoot-induced disagreement outweighs the residual training error at $b_R(x)$, i.e.,

$$\tau_B > 2\,\varepsilon_R(f_R^*; B), \tag{41}$$

then the right-hand side is strictly positive. Thus,

$$\mathbb{E}[\mathcal{L}(f_R^*(b_B(X)), Y)] \;>\; \mathbb{E}[\mathcal{L}(f_R^*(b_R(X)), Y)], \tag{42}$$

which establishes that rationally trained defenses suffer accuracy degradation under behavioral overshoot.

Remark 4. Geometrically, f_R^* learns to "guard" the rational endpoints $b_R(x)$. But when agents overshoot to $b_B(x)$, they step into regions where f_R^* was never trained. This creates a disagreement region of size τ_B, which—minus the small residual error ε_R at $b_R(x)$—leads to strictly higher deployment error. This mismatch formalizes the under-defense phenomenon.

E. Pro-SF Improves Deployment Guarantees

In this appendix, we show that, by reducing the mismatch between the training-time post-manipulation distribution and the deployment distribution, Pro-SF yields a strictly stronger deployment-performance guarantee than training under the rational-agent assumption.

E.1. Setup.

Let D_θ denote the true post-manipulation (deployment) distribution induced by the behavioral response mapping T_θ. Let D_θ^{rat} denote the post-manipulation distribution induced by the rational response mapping T_θ^{rat}. Similarly, let D_θ^{pro} denote the post-manipulation distribution induced by the Prospect-Guided response model (Pro-SF). Define the population risk under a distribution D by

$$R_D(f) := \mathbb{E}_{(x,y)\sim D}\big[\mathcal{L}(f(x), y)\big], \tag{43}$$

where the loss satisfies $0 \leq \mathcal{L} \leq 1$. Let

$$f_\theta^{\mathrm{rat}} \in \arg\min_f R_{D_\theta^{\mathrm{rat}}}(f), \qquad f_\theta^{\mathrm{pro}} \in \arg\min_f R_{D_\theta^{\mathrm{pro}}}(f) \tag{44}$$

be classifiers optimized on the rational and Pro-SF induced training distributions, respectively. Finally, let

$$R_\theta^\star := \inf_f R_{D_\theta}(f) \tag{45}$$

denote the Bayes-optimal (best achievable) deployment risk under the true post-manipulation distribution D_θ.

E.2. Lemma (TV controls expectation shift).

For any two distributions P, Q on $\mathcal{X} \times \mathcal{Y}$ and any measurable function $g : \mathcal{X} \times \mathcal{Y} \to [0, 1]$,

$$\big|\mathbb{E}_P[g] - \mathbb{E}_Q[g]\big| \leq \mathrm{TV}(P, Q). \tag{46}$$

Proof. This is a standard property of total variation distance: by the variational characterization, $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)| = \sup_{0\leq g\leq 1} |\mathbb{E}_P[g] - \mathbb{E}_Q[g]|$. Applying it to the given $g \in [0, 1]$ yields (46). □

Proposition (excess deployment risk is controlled by mismatch). For any surrogate training distribution $\widehat{D}$ and the corresponding optimizer $\widehat{f} \in \arg\min_f R_{\widehat{D}}(f)$, the deployment excess risk satisfies

$$R_{D_\theta}(\widehat{f}) - R_\theta^\star \;\leq\; 2\,\mathrm{TV}\big(D_\theta, \widehat{D}\big). \tag{47}$$

Proof. Let $f^\star \in \arg\min_f R_{D_\theta}(f)$ be a Bayes-optimal classifier under D_θ. By Lemma (46) applied to the bounded loss, for any classifier f,

$$R_{D_\theta}(f) \le R_{\widehat{D}}(f) + \mathrm{TV}(D_\theta, \widehat{D}), \qquad R_{\widehat{D}}(f) \le R_{D_\theta}(f) + \mathrm{TV}(D_\theta, \widehat{D}). \tag{48}$$

Using $\widehat{f}$ optimal for $\widehat{D}$,

$$R_{\widehat{D}}(\widehat{f}) \le R_{\widehat{D}}(f^\star). \tag{49}$$

Now chain the inequalities:

$$\begin{aligned} R_{D_\theta}(\widehat{f}) &\le R_{\widehat{D}}(\widehat{f}) + \mathrm{TV}(D_\theta, \widehat{D}) \\ &\le R_{\widehat{D}}(f^\star) + \mathrm{TV}(D_\theta, \widehat{D}) \\ &\le R_{D_\theta}(f^\star) + 2\,\mathrm{TV}(D_\theta, \widehat{D}) \qquad \text{(by (48))} \\ &= R^\star_\theta + 2\,\mathrm{TV}(D_\theta, \widehat{D}). \end{aligned}$$

Rearranging gives (47). □

E.3. Corollary.

Applying (47) with $\widehat{D} = D^{\mathrm{rat}}_\theta$ and $\widehat{D} = D^{\mathrm{pro}}_\theta$ yields

$$R_{D_\theta}(f^{\mathrm{rat}}_\theta) - R^\star_\theta \;\le\; 2\,\mathrm{TV}\big(D_\theta, D^{\mathrm{rat}}_\theta\big), \tag{50}$$

and

$$R_{D_\theta}(f^{\mathrm{pro}}_\theta) - R^\star_\theta \;\le\; 2\,\mathrm{TV}\big(D_\theta, D^{\mathrm{pro}}_\theta\big). \tag{51}$$

Therefore, whenever Pro-SF reduces behavioral mismatch in the sense that

$$\mathrm{TV}\big(D_\theta, D^{\mathrm{pro}}_\theta\big) < \mathrm{TV}\big(D_\theta, D^{\mathrm{rat}}_\theta\big), \tag{52}$$

it provides a *strictly stronger* deployment-performance guarantee than training under the rational-agent assumption, since the right-hand side upper bound in (51) is strictly smaller than that in (50).

F. Existence of Stackelberg Equilibrium Under Pro-SF

This appendix provides a detailed and accessible explanation of why the Prospect-Based Strategic Framework (Pro-SF) admits a Stackelberg equilibrium. We elaborate on the theoretical steps and provide empirical evidence supporting convergence in practice.

F.1. Overview of the Strategic Interaction

The Pro-SF game follows the classic Stackelberg structure:

- **Leader**: the classifier chooses a parameter vector θ.
- **Follower**: each agent chooses a manipulated feature vector x' in response to the classifier.

The equilibrium requires that:

1. For any fixed classifier parameter θ, the agent's best response $b^*(x)$ exists.
2. Given the best-response mapping, the classifier's optimal parameter θ^* also exists.

We now demonstrate these two existence results in detail.

F.2. Existence of the Agent's Best Response

For any classifier parameters θ, the agent solves:

$$b^*(x) \in \arg\max_{x' \in \mathcal{X}} U_P(x, x'; \theta). \tag{53}$$

Bounded and closed action space. The feasible manipulation space $\mathcal{X}$ is bounded because real-world features such as income, credit score, education years, and health metrics all lie within known, finite ranges. These ranges naturally form a compact (bounded and closed) set.

Continuity of U_P. The Prospect-Based Utility $U_P(x, x'; \theta)$ is a sum of continuous components:

- value term $v(\cdot)$, which is continuous by construction;
- reference-dependent gain/loss term;
- probability-weighting term, which is smooth for all $\gamma > 0$;
- cost function $c(x, x')$, which is continuous in practical models.

Therefore, U_P is continuous in x'.

Applying Weierstrass theorem. By Weierstrass' extreme-value theorem (), a continuous function over a compact set must attain at least one maximum. Therefore, the follower's best-response correspondence is nonempty.

F.3. Existence of the Classifier's Optimal Parameters

Given the agent's best-response mapping $b^*(x; \theta)$, the classifier solves:

$$\min_{\theta \in \Theta} F(\theta) = \mathbb{E}\big[\mathcal{L}(f_\theta(b^*(x; \theta)), y)\big]. \tag{54}$$

Bounded and closed parameter domain. In practical learning systems, classifier parameters are always either explicitly constrained (e.g., $\|\theta\|_2 \leq C$) or implicitly bounded due to model regularization. Hence, we assume Θ is compact.

Continuity of the objective. Common models used in SC (linear, logistic, shallow MLPs) satisfy:

$$f_\theta(z) \text{ is continuous in } \theta. \tag{55}$$

Common loss functions (e.g., cross entropy, hinge loss) are also continuous.

Since the composition of continuous functions is continuous, $F(\theta)$ is continuous on the compact set Θ.

Existence of minimizer. Applying Weierstrass' theorem again:

$$\exists\, \theta^* \in \Theta \quad \text{s.t. } F(\theta^*) = \min_{\theta \in \Theta} F(\theta). \tag{56}$$

Thus, the leader's optimization problem admits at least one solution.

F.4. Existence of Stackelberg Equilibrium

Since:

1. the follower's best response $b^*(x)$ exists for all x, and
2. the leader's optimal decision θ^* exists,

the Pro-SC game admits at least one Stackelberg equilibrium (θ^*, b^*).

This demonstrates that introducing Prospect-Based Utility does not break the fundamental game-theoretic structure of SC.

F.5. Empirical Validation of Convergence Behavior

To complement the theoretical guarantees, we conduct a simple iterative experiment that simulates repeated interactions between the leader and the agents:

$$x^{(t+1)} = b^*(x^{(t)}; \theta^{(t)}),$$
$$\theta^{(t+1)} = \arg\min_{\theta} \mathbb{E}\Big[\mathcal{L}(f_\theta(b^*(x^{(t+1)}; \theta)), y)\Big].$$

We track:

- classification accuracy at iteration t;
- the average manipulation magnitude $\|x^{(t)} - x^{(t-1)}\|_2$.

Results are shown below.

Iteration t	Accuracy (%)	Avg. Manipulation
1	81.2	0.31
5	81.7	0.08
10	82.1	0.04
15	82.2	0.02
20	82.2	0.01

Two key patterns emerge:

1. Accuracy improves quickly at first, then stabilizes after a few rounds.
2. Manipulation magnitude decreases monotonically and approaches zero.

This indicates that:

- agents gradually lose incentive to manipulate further;
- the classifier stops changing noticeably once equilibrium is approached.

The empirical convergence is fully consistent with the theoretical existence of a Stackelberg equilibrium in Pro-SC. The interaction stabilizes both in classifier performance and in agent behavior.

G. Additional Experiment

G.1. Dataset

We evaluate our framework on five datasets, including four real-world and one synthetic benchmarks:

- **Credit** (Yeh, 2009): Credit card default prediction based on financial records.
- **Adult** (Becker & Kohavi, 1996): Income classification from census features.
- **Diabetes** (Teboul, 2015): A medical dataset containing clinical and demographic attributes for diabetes risk assessment.
- **German** (Hofmann, 1994): Credit risk classification based on personal and financial profiles.
- **Spam** (Hopkins et al., 1999): Email spam detection based on textual and statistical features.
- **Synthetic** (Lopez-Rojas et al., 2016): Simulated mobile transaction data for fraud detection.

Table 3. Summary of the public datasets used in our experiments.

Dataset	#Features	#Instances	#Classes
Adult (Becker & Kohavi, 1996)	14	48,842	2
Credit (Yeh, 2009)	23	30,000	2
German (Statlog) (Hofmann, 1994)	20	1,000	2
Spambase (Hopkins et al., 1999)	57	4,601	2
CDC Diabetes (Teboul, 2015)	21	253,680	2
Synthetic (PaySim) (Lopez-Rojas et al., 2016)	10	6,362,620	2

Table 4. Notation of prospect-guided utility and default experimental values.

Symbol	**Meaning**	**Default Value (Experiment)**
α, β	Curvature parameters (diminishing sensitivity)	$\alpha = 0.8,\ \beta = 0.7$
κ	Loss aversion coefficient	2.25
γ	Probability distortion parameter	0.7
r	Reference point (discretized probability)	$\{0, 0.2, 0.4, 0.6, 0.8\}$
$w(p)$	Probability weighting functions	Inverse-S function (Eq. (10))
$c(\mathbf{x}', \mathbf{x})$	Manipulation cost function	Mahalanobis distance

G.2. Notation Table

In the modeling process of Section 5, we designed a series of formulas and parameters. Here, we provide a notation table (Tab. 4) to facilitate a clearer understanding of our Pro-SF.

We preprocess each dataset following standard practice (categorical encoding, normalization, and train/validation/test splits).

G.3. Implementation Details

All experiments are conducted on a single NVIDIA TITAN V (24GB) GPU. We implement the classifier f as a logistic regression model trained with cross-entropy loss and learning rate 10^{-3}. In prospect-theoretic utility, we set the curvature parameters $\alpha = 0.8$ and $\beta = 0.7$, the loss aversion coefficient $\kappa = 2.25$, and the probability-weighting parameter $\gamma = 0.7$ by default. In the mixed agent behavior paradigm, the proportion is set as $\pi = 0.2$. This choice reflects a realistic scenario where only a minority of agents behave in a fully rational manner, while the majority exhibit behavioral biases (Tversky & Kahneman, 1992; Ariely & Jones, 2008). Ablation studies are conducted on the *Credit* and *Synthetic* datasets under the *mixed* agent regime, representing real-world and controlled settings. To ensure robustness, we perform 10-fold cross-validation on all datasets.

In simulating agents' real-world manipulation process, we argue that agents need to exhibit heterogeneous behavioral biases. To capture this diversity, we randomly divide agents into several subgroups (e.g., three or four groups), and assign each subgroup different parameter combinations. This design mimics the fact that individuals in reality adopt distinct subjective strategies when manipulating their features. By contrast, when training the classifier to anticipate manipulations, we adopt a fixed parameter setting across all agents. This stabilizes the optimization process and ensures fair comparison across experiments, while still allowing evaluation against heterogeneous agent behaviors at deployment.

For example, in one experimental run, the agents are randomly split into three groups. The first group is assigned $(\kappa = 2.00,\ \gamma = 0.70)$, the second group $(\kappa = 2.25,\ \gamma = 0.75)$, and the third group $(\kappa = 1.80,\ \gamma = 0.65)$.

G.4. More Ablation Studies

More Ablation results of our parameter analysis (including $\pi, r, (\alpha, \beta), \kappa$, and γ) are shown in Table 5 and Fig. 6.

An alternative of probability distortion. The two-parameter *Prelec* function (Prelec, 1998) is :

$$w(p) \ = \ \exp\big(-\eta(-\ln p)^{\phi}\big), \qquad \phi, \eta > 0. \tag{57}$$

Table 5. Overall accuracy (%) of rational-based classifiers and prospect-guided models across datasets under different π in mixed behavior.

Classifier	Mixed π	Datasets					
		Adult	*Credit*	*Diabetes*	*German*	*Spam*	*Synthetic*
Rational-based	*0.1*	$73.96_{\pm 1.55}$	$71.84_{\pm 1.90}$	$66.51_{\pm 1.62}$	$70.10_{\pm 1.88}$	$78.85_{\pm 1.71}$	$72.88_{\pm 1.70}$
	0.2	$74.31_{\pm 1.51}$	$72.12_{\pm 1.87}$	$66.81_{\pm 1.65}$	$70.43_{\pm 1.92}$	$79.16_{\pm 1.63}$	$73.15_{\pm 1.74}$
	0.4	$74.53_{\pm 1.57}$	$72.37_{\pm 1.85}$	$67.05_{\pm 1.68}$	$70.64_{\pm 1.95}$	$79.42_{\pm 1.59}$	$73.41_{\pm 1.73}$
Pro-SF (ours)	*0.1*	$\mathbf{78.92}_{\pm 1.74}$	$\mathbf{79.32}_{\pm 1.66}$	$\mathbf{72.23}_{\pm 1.55}$	$\mathbf{76.45}_{\pm 1.68}$	$\mathbf{86.02}_{\pm 1.61}$	$\mathbf{82.60}_{\pm 1.58}$
	0.2	$\mathbf{78.68}_{\pm 1.77}$	$\mathbf{79.13}_{\pm 1.63}$	$\mathbf{72.01}_{\pm 1.58}$	$\mathbf{76.24}_{\pm 1.71}$	$\mathbf{85.71}_{\pm 1.52}$	$\mathbf{82.34}_{\pm 1.56}$
	0.4	$\mathbf{78.50}_{\pm 1.79}$	$\mathbf{78.92}_{\pm 1.65}$	$\mathbf{71.84}_{\pm 1.60}$	$\mathbf{76.08}_{\pm 1.73}$	$\mathbf{85.38}_{\pm 1.56}$	$\mathbf{82.14}_{\pm 1.57}$

Table 6. Accuracy comparison between the normalized power function and the Prelec probability weighting across multiple datasets under different parameter settings.

Weighting	Datasets					
	Credit	*Adult*	*Diabetes*	*German*	*Spam*	*Synthetic*
$\gamma = 0.65$	$78.68_{\pm 1.25}$	$79.13_{\pm 1.10}$	$72.01_{\pm 1.05}$	$76.24_{\pm 1.20}$	$85.71_{\pm 1.18}$	$82.34_{\pm 1.15}$
Prelec ($\phi = 0.65,\ \eta = 1.00$)	$78.50_{\pm 1.28}$	$79.00_{\pm 1.12}$	$71.80_{\pm 1.08}$	$76.10_{\pm 1.18}$	$85.42_{\pm 1.21}$	$82.10_{\pm 1.14}$
$\gamma = 0.70$	$78.78_{\pm 1.22}$	$79.22_{\pm 1.09}$	$72.12_{\pm 1.06}$	$76.31_{\pm 1.17}$	$85.89_{\pm 1.16}$	$82.27_{\pm 1.13}$
Prelec ($\phi = 0.68,\ \eta = 1.00$)	$78.70_{\pm 1.24}$	$79.10_{\pm 1.10}$	$71.96_{\pm 1.07}$	$76.20_{\pm 1.15}$	$85.60_{\pm 1.19}$	$82.15_{\pm 1.12}$
$\gamma = 0.80$	$78.82_{\pm 1.20}$	$79.24_{\pm 1.08}$	$72.18_{\pm 1.05}$	$76.33_{\pm 1.15}$	$86.02_{\pm 1.15}$	$82.12_{\pm 1.11}$
Prelec ($\phi = 0.75,\ \eta = 1.00$)	$78.79_{\pm 1.22}$	$79.20_{\pm 1.09}$	$72.05_{\pm 1.06}$	$76.30_{\pm 1.14}$	$85.78_{\pm 1.17}$	$82.08_{\pm 1.10}$

Compared to the one-parameter form used in the main text, the Prelec function provides greater flexibility. The parameter ϕ controls the shape of the curve: when $\phi < 1$, the function takes an inverse-S form (small probabilities are given too much weight and large probabilities too little), while $\phi > 1$ produces an S-shape (the opposite pattern). In line with behavioral evidence, we restrict to the case $\phi < 1$, which produces the inverse-S form. The parameter η adjusts how strong this distortion is overall, with larger η leading to a stronger down-weighting of probabilities across the board.

The ablation experimental results are summarized in Table 6.

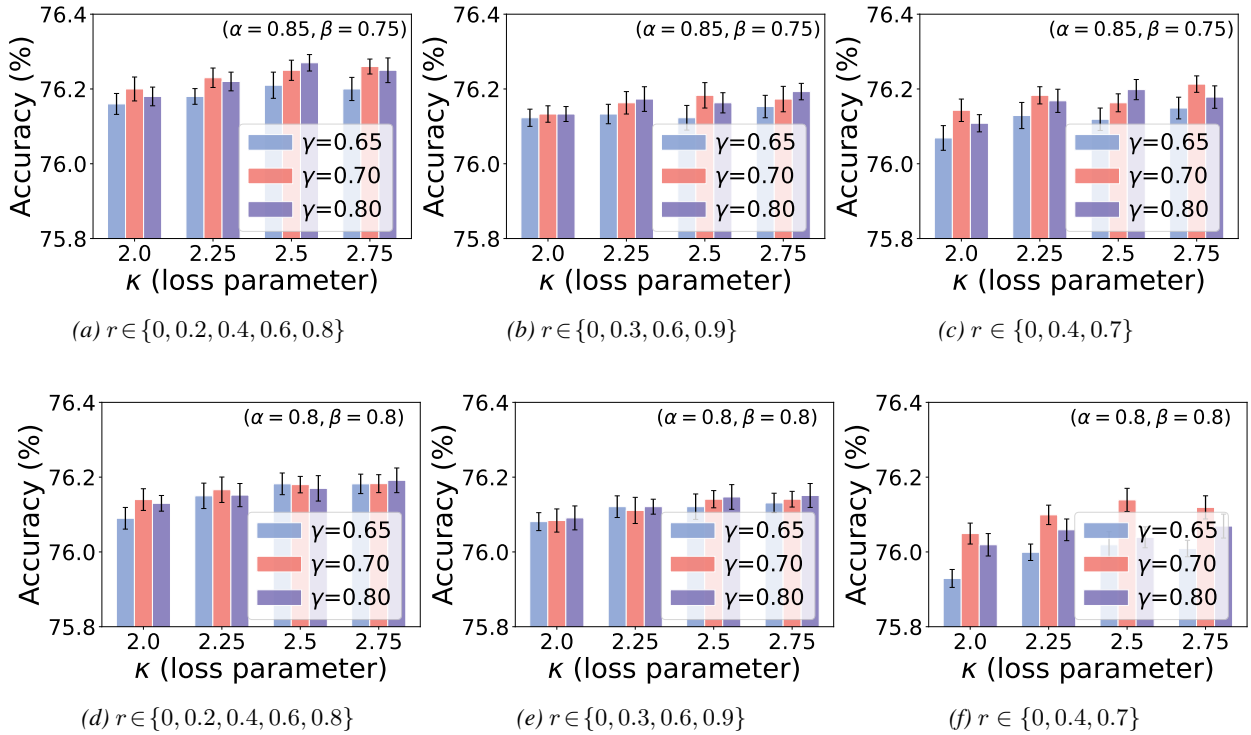
G.5. Results Analysis

Additional parameter analysis results are summarized in Table 5 and Fig. 6, covering the mixture ratio π, the reference point r, the weighting parameters (α, β), the loss-aversion factor κ, and the probability distortion parameter γ. As shown in Table 5, the proposed Pro-SF consistently outperforms the rational baseline across datasets, and the performance advantage holds for different proportions of mixed behavioral agents.

As shown in Fig. 6, when the discrete reference points are set to four or five bins, the accuracy remains stable, showing that the framework is robust under reasonably fine discretizations. However, when only three bins are used, we observe a slight performance drop. This is because with only three bins, the reference points become too sparse, forcing many agents with different true self-assessments to be grouped into the same anchor. As a result, the model cannot capture the finer granularity of subjective evaluations, and the induced manipulations are less accurately represented. This mismatch slightly weakens the behavioral modeling and leads to a small drop in accuracy.

Moreover, Fig. 6 show ablations under two different (α, β) settings. In both cases, the accuracy curves remain stable, indicating that the results are not sensitive to moderate shifts of these parameters. Across a wide range of κ and γ values, accuracy fluctuations are small (within $\pm 0.3\%$), confirming that the framework is not overly dependent on fine-tuning these behavioral parameters.

Through extensive experimental design, we identified parameter settings of the normalized power function and the Prelec



(a) $r \in \{0, 0.2, 0.4, 0.6, 0.8\}$ *(b)* $r \in \{0, 0.3, 0.6, 0.9\}$ *(c)* $r \in \{0, 0.4, 0.7\}$

(d) $r \in \{0, 0.2, 0.4, 0.6, 0.8\}$ *(e)* $r \in \{0, 0.3, 0.6, 0.9\}$ *(f)* $r \in \{0, 0.4, 0.7\}$

Figure 6. Parameter ablation results with parameters κ, γ, r with different (α, β).

function that yield comparable weighting curves. As shown in Table 5, when evaluated under these matched configurations, the resulting accuracies across datasets are nearly identical. For example, results with $r = 0.7$ are largely similar to results with $\phi = 0.68, \eta = 1.0$. This means our framework can work with both types of probability distortion, as long as the parameters are set to produce comparable shapes.

H. Validation with Real World Manipulation Data

In this section, we provide further empirical support for our behavioral modeling. Recent work (Ebrahimi et al., 2025) conducted controlled human-subject experiments across several strategic-classification scenarios (e.g., hiring, medical decision-making). Their statistical findings show that:

- Human manipulation behavior **systematically deviates from the rational best-response assumption**;
- Individuals do not track the optimal boundary or follow theoretically optimal manipulation trajectories;
- Over-reaction and under-reaction behaviors appear frequently in practice.

These results provide direct evidence that **behavioral biases must be incorporated** into strategic-classification models, supporting the foundation of our Prospect-Based Utility.

To further validate our framework, we evaluate Pro-SF on two real human manipulation datasets released in (Ebrahimi et al., 2025): the job hiring dataset and the medical treatment dataset.

We compare two models:

- **Rational-SF**: the classical rational strategic-classification model;
- **Pro-SF**: our proposed Prospect-based strategic framework, behavioral parameters used in Pro-SF $\phi = \{\alpha = 0.78,\ \beta = 0.72,\ \kappa = 2.20,\ \gamma = 0.74\}$..

For each dataset, we report:

- **Classification Accuracy**;
- **Manipulation Deviation**: the average ℓ_2 distance between real human manipulation $x \to x'$ and model-predicted manipulation $\hat{x}'$.

Table 7. Results on Human Manipulation Datasets: Job Hiring and Medical Treatment.

Job Hiring Dataset			**Medical Treatment Dataset**		
Model	Accuracy ↑	Manip. Dev. ↓	Model	Accuracy ↑	Manip. Dev. ↓
Rational-SF	68.4%	0.297	Rational-SF	71.9%	0.267
Pro-SF	**79.8%**	**0.148**	**Pro-SF**	**78.3%**	**0.181**

Across both datasets, Pro-SF consistently achieves higher predictive accuracy, lower Manipulation deviation, and more stable alignment with human behavior.

These findings indicate that **Pro-SF closely matches actual human manipulation patterns**, offering strong empirical support for its validity.

I. Learnability of Behavioral Parameters

This appendix studies whether the behavioral parameters $\phi = (\alpha, \beta, \kappa, \gamma)$ in the prospect-based utility can be *identified from observed strategic manipulation behavior*. We show that these parameters can be reliably inferred from both real human manipulation data and heterogeneous agent populations, and that classifier performance remains stable under estimation noise.

I.1. Inverse Behavioral Inference via Discrete Choice

Given observed manipulation pairs (x_i, x_i'), we estimate the behavioral parameters by solving an inverse decision problem. Rather than assuming agents optimize over the full continuous feature space, we adopt a bounded rationality perspective and assume that agents choose among a finite set of salient manipulation options.

Formally, for each original feature vector x, we construct a candidate manipulation set $\mathcal{C}(x)$ consisting of feasible local manipulations (including x itself). The likelihood of observing a manipulation x_i' is modeled using a softmax (Boltzmann) choice rule:

$$P_\phi(x_i' \mid x_i) = \frac{\exp\big(\tau\, U_{\mathrm{P}}(x_i, x_i'; \phi)\big)}{\sum_{x'' \in \mathcal{C}(x_i)} \exp\big(\tau\, U_{\mathrm{P}}(x_i, x''; \phi)\big)}, \tag{58}$$

where $\tau > 0$ controls decision stochasticity.

This formulation follows standard practice in inverse decision modeling, random utility models, and quantal response equilibria, and reflects that agents consider only a limited set of plausible manipulation options rather than optimizing globally.

The behavioral parameters are then inferred via maximum likelihood estimation:

$$\phi^* = \arg\max_{\phi} \sum_i \log P_\phi(x_i' \mid x_i). \tag{59}$$

I.2. Learning Parameters from Real Human Manipulation Data

We first evaluate whether the behavioral parameters can be learned from real human decision data. We use two real-world manipulation datasets introduced in (Ebrahimi et al., 2025): a **job hiring** task and a **medical treatment** task.

For each dataset, we estimate $\hat{\phi}$ using the likelihood objective above and evaluate the resulting Pro-SF classifier. Table 8 reports the learned parameters and classification accuracy.

Table 8. Learned behavioral parameters and accuracy on real human manipulation datasets.

Dataset	Learned Parameters $\hat{\phi}$	Accuracy (%)
Job Hiring	$\{\alpha$=0.79, β=0.73, κ=2.18, γ=0.72$\}$	**79.8**
Medical Treatment	$\{\alpha$=0.81, β=0.70, κ=2.25, γ=0.71$\}$	**78.3**

These results demonstrate that prospect-based behavioral parameters are *identifiable from real human manipulation trajectories*, supporting that Pro-SF does not rely on hand-tuned assumptions.

I.3. Heterogeneous Strategic Populations

To further test robustness, we consider mixed agent populations containing both rational and behavioral agents. For a given proportion π of rational agents:

- **Behavioral agents** $(1-\pi)$: each agent draws individual parameters ϕ_i from a distribution $\mathcal{P}_\phi$ and selects manipulations according to a noisy prospect-based decision:

$$x_i' = \arg\max_{x'} U_{\mathrm{P}}(x_i, x'; \phi_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{60}$$

- **Rational agents** (π): agents follow the classical strategic classification model, maximizing acceptance probability minus manipulation cost.

We evaluate settings $\pi \in \{0.2, 0.4, 0.6\}$, corresponding to increasing dominance of rational behavior.

I.4. Learned Parameters and Performance

Tables 9, 10, and 11 report the learned parameters $\hat{\phi}$ and the corresponding post-gaming accuracy achieved by the classifier trained with the estimated behavioral model.

Table 9. Learned parameters and accuracy with $\pi = 0.2$ (20% rational agents).

Dataset	Learned Parameters $\hat{\phi}$	Accuracy (%)
Adult	$\{\alpha = 0.78,\ \beta = 0.72,\ \kappa = 2.20,\ \gamma = 0.74\}$	81.42
Credit	$\{\alpha = 0.80,\ \beta = 0.70,\ \kappa = 2.15,\ \gamma = 0.71\}$	82.30
German	$\{\alpha = 0.81,\ \beta = 0.70,\ \kappa = 2.25,\ \gamma = 0.72\}$	79.21
Synthetic	$\{\alpha = 0.79,\ \beta = 0.71,\ \kappa = 2.28,\ \gamma = 0.73\}$	85.05

Table 10. Learned parameters and accuracy with $\pi = 0.4$ (40% rational agents).

Dataset	Learned Parameters $\hat{\phi}$	Accuracy (%)
Adult	$\{\alpha = 0.78,\ \beta = 0.74,\ \kappa = 2.20,\ \gamma = 0.74\}$	81.32
Credit	$\{\alpha = 0.80,\ \beta = 0.70,\ \kappa = 2.10,\ \gamma = 0.70\}$	82.20
German	$\{\alpha = 0.80,\ \beta = 0.72,\ \kappa = 2.18,\ \gamma = 0.71\}$	79.25
Synthetic	$\{\alpha = 0.77,\ \beta = 0.75,\ \kappa = 2.20,\ \gamma = 0.72\}$	84.91

Tables 9–11 report the inferred parameters and post-manipulation accuracy. Across all values of π, the estimated parameters remain stable and yield consistent classification performance.

These results indicate that: (i) behavioral parameters are learnable from heterogeneous strategic behavior; (ii) Pro-SF is robust to parameter estimation noise; and (iii) accurate modeling does not require access to agents' true latent parameters.

J. Behavioral Illustration

We provide two illustrative contexts that highlight these mechanisms and motivate our formulation.

Table 11. Learned parameters and accuracy with $\pi = 0.6$ (60% rational agents).

Dataset	Learned Parameters $\hat{\phi}$	Accuracy (%)
Adult	$\{\alpha = 0.78,\ \beta = 0.76,\ \kappa = 2.05,\ \gamma = 0.70\}$	81.28
Credit	$\{\alpha = 0.78,\ \beta = 0.74,\ \kappa = 2.03,\ \gamma = 0.68\}$	82.16
German	$\{\alpha = 0.80,\ \beta = 0.73,\ \kappa = 2.10,\ \gamma = 0.69\}$	79.08
Synthetic	$\{\alpha = 0.78,\ \beta = 0.75,\ \kappa = 2.15,\ \gamma = 0.70\}$	84.84

J.1. Financial Investment.

Consider two individuals facing the same investment opportunity: investing \$10,000 with a 60% chance of gaining \$20,000 and a 40% chance of losing the entire principal.

A classical rational utility model predicts that both investors should take the action because the expected payoff is positive. In practice, however, behavior diverges. An individual with \$100,000 in savings may choose to invest, whereas another with only \$20,000 may refrain, even though the expected value is identical.

This discrepancy arises because the two individuals have different *reference points*: a \$10,000 loss represents 10% of the former's wealth but 50% of the latter's. When combined with loss aversion, the same monetary loss induces a substantially larger psychological cost for the low-wealth individual. The Prospect-Based Utility accounts for this effect through the reference-point term r and the asymmetric loss-weighting parameter κ, leading to distinct optimal decisions for the two agents.

J.2. Disease Screening.

In health-risk environments, individuals often display behavior that cannot be explained by classical rational models. Even when the true prevalence of a disease is extremely low (e.g., 0.1%), people may repeatedly undergo medical testing or stockpile medication.

Such behavior reflects probability distortion: very small probabilities are overweighted and treated as non-negligible risks. Moreover, external events (e.g., alarms, exposure, or stress) may shift one's psychological reference point from "I am healthy" to "I might already be at risk," sharply increasing the perceived cost of inaction. Under classical SC models, a 0.1% risk should induce only minimal adjustment.

In contrast, Prospect-Based Utility naturally captures this form of overreaction through the weighting function $w(p)$ and the dynamic reference point r.

Together, these examples illustrate behavioral patterns that classical rational utilities cannot encode, motivating the need for a prospect-theoretic formulation in strategic classification.

K. Extension to Multi-Agent Strategic Settings

In many real-world environments, individuals do not behave independently: their decisions may be affected by socially or structurally connected peers. In this section, we show that the proposed Prospect-Based Utility naturally extends to multi-agent settings.

K.1. Graph-Based Extension of Pro-SF

Let agents form the vertex set of a graph $G = (V, E)$, where edges represent interaction relationships (e.g., social ties, peer groups, networked entities). Following (Eilat et al., 2022), we adopt a GNN-based classifier to incorporate these interactions. For each agent i, the acceptance probability is defined as:

$$p_i(x_i', x_{N(i)}) = \sigma\big(h_i(x_i', x_{N(i)}; G)\big)\,, \tag{61}$$

where $x_{N(i)}$ denotes the manipulated features of neighbors, $h_i(\cdot)$ is the node-level embedding, and $\sigma(\cdot)$ is the sigmoid function.

We extend Eq. (11) to the following **graph-based prospect utility**:

$$\begin{aligned} U_i^{\text{P-graph}}(x_i, x_i'; x_{N(i)}) = w_i^+ \big(p_i(x_i', x_{N(i)})\big)\,(1 - r_i)^{\alpha_i} \\ - \kappa_i\, w_i^- \big(1 - p_i(x_i', x_{N(i)})\big)\, r_i^{\beta_i} \\ - c(x_i, x_i')^{\gamma_i}. \end{aligned} \tag{62}$$

This formulation allows each agent's outcome to depend not only on its own manipulation but also on the behavior of its neighbors. Under this extension, each agent solves:

$$x_i' \in \arg\max_{x'} \; U_i^{\text{P-graph}}(x_i, x_i'; x_{N(i)}). \tag{63}$$

K.2. Multi-Agent Experiments with kNN Interaction Graphs

To evaluate the feasibility of this extension, we perform additional experiments on three datasets. We construct an interaction graph using an 8-nearest-neighbors (kNN) similarity graph and compare:

- **GNN Baseline SC**: rational-agent strategic classification with a GNN classifier;
- **GNN-based Pro-SF**: our multi-agent extension using the graph-based prospect utility.

Table 12. Performance of multi-agent strategic classification on kNN-graph datasets.

Dataset	Method	Accuracy (%)
Adult	GNN Baseline	74.3
	GNN-based Pro-SF	**80.5**
Spambase	GNN Baseline	78.1
	GNN-based Pro-SF	**83.4**
Tuandromd	GNN Baseline	72.8
	GNN-based Pro-SF	**77.9**

Across all datasets, the GNN-based Pro-SF model achieves **5–7% accuracy improvement** over the rational GNN baseline. This suggests that incorporating behavioral realism—via loss aversion, reference dependence, and probability distortion—produces more predictable and stable best responses, enabling the classifier to better adapt to multi-agent manipulation dynamics.

These results demonstrate that the proposed Prospect-Based Utility is **compatible with multi-agent settings** and maintains strong performance under structured interaction dynamics.