Partial Fairness Awareness: Belief-Guided Strategic Mechanism for Strategic Agents

Xinpeng Lv[1], Chunyuan Zheng[2], Yunxin Mao[1], Renzhe Xu[3], Hao Zou[4], Shanzhi Gu[1], Liyang Xu[1], Huan Chen[1], Yuanlong Chen[5], Wenjing Yang[1], Haotian Wang[1*]

[1]National University of Defense Technology, Changsha, China
[2]Peking University, Beijing, China
[3]Shanghai University of Finance and Economics, Shanghai, China
[4]ZGC laboratory, Beijing, China
[5]Faculty of Computing, Harbin Institute of Technology, Harbin, China
{lvxinpeng, maoyunxin, wanghaotian13}@nudt.edu.cn

Abstract

Strategic machine learning investigates scenarios where agents manipulate their features to receive favorable decisions from predictive models. To address fairness concerns intrinsic to strategic classification, recent work has introduced group-specific fairness constraints. However, current fairness-aware approaches face a fundamental dilemma in the issue of fairness exposure: making these constraints public enables strategic manipulation and can lead to fairness reversal, while keeping them hidden may reduce social welfare and discourage genuine improvement. To fill this gap, we subsequently propose the problem of ***P**artial **F**airness **A**wareness* (PFA), as our theoretical analysis informs that such a dilemma can be mitigated by releasing the candidate set of fairness constraints and concealing the grounding constraint. To be specific, we introduce a **belief-guided strategic mechanism**, wherein agents iteratively interact with the decision system and maintain a belief distribution over the candidate set of fairness constraints. This belief-guided process enables agents, through iterative interaction and feedback, to update their belief distribution over the candidate set, thereby gradually aligning their belief with the grounding fairness constraint employed by the system. Extensive experiments on real-world and synthetic datasets demonstrate that PFA achieves lower group fairness gaps, higher acceptance of truly qualified individuals, and more stable outcomes compared to fully public or private fairness regimes.

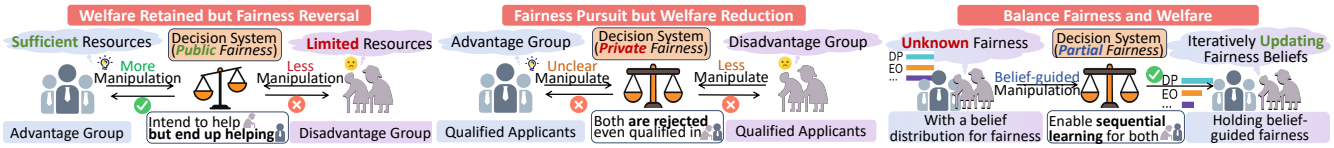
Introduction

Machine learning models are increasingly deployed in decision-making domains such as hiring (Sánchez-Monedero, Dencik, and Edwards 2020), credit scoring (Jagtiani and Lemieux 2019), and college admissions (Kučak, Juričić, and DJambić 2018). In these scenarios, individuals (agents) often engage in strategic manipulation of their features to obtain favorable outcomes. As noted by Goodhart's Law (Strathern 1997), "*When a measure becomes a target, it ceases to be a good measure*," such gaming behaviors can undermine the reliability of decision models. For example, a loan applicant might temporarily inflate their reported income to appear more creditworthy. To ensure robustness towards the strategic manipulations, the strategic classification (SC) framework (Hardt et al. 2016) has been developed to model Stackelberg-style interaction between model designers and strategic agents (Ghalme et al. 2021; Singh and Kulkarni 2024; Chen, Liu, and Podimata 2020), aiming to maintain predictive accuracy under adversarial conditions.

Despite the remarkable progress of the SC framework in robustness, practical deployment also necessitates careful consideration of fairness for sensitive attributes or historically disadvantaged groups. In particular, a model resistant to gaming may still systematically disadvantage certain groups, i.e., exhibiting *unfairness*. For example, even if a hiring platform exhibits robustness to manipulation, historical biases embedded in the data might lead to unfair judgment towards female applicants, reflecting and reinforcing social inequities (Zemel et al. 2013; Zhang et al. 2022). To address the *fairness concern in strategic-robust decision models*, a growing body of work has incorporated group fairness constraints, such as demographic parity (Zemel et al. 2013), equality of opportunity (Roemer and Trannoy 2015), and predictive parity (Dieterich, Mendoza, and Brennan 2016), into strategic classification, typically through fairness-aware decision rules adapted for different groups (Zhang et al. 2022; Shimao et al. 2025; Zhang et al. 2022).

However, most concurrent efforts focus on designing more effective fairness mechanisms that are fully public to agents, while overlooking a more fundamental consideration: *whether and how fairness constraints should be exposed to agents*. In practice, different exposure of the fairness constraints results in divergent subsequent fairness mechanism design, leading to different algorithmic choices and social welfare variation. On the one hand, in the case of fully public constraints, previous theories already point out the unexpected phenomenon of *fairness reversal* (Estornell et al. 2023b) (public fairness in Figure 1(a)), where advantaged individuals manipulate based on the exposed fairness constraints. Hence, the reverse effect of fairness occurs, as the intended benefits for disadvantaged groups are conversely further undermined. On the other hand, when the fairness constraints are fully concealed (private fairness in Figure 1(b)), our theoretical analysis proves that the social

*Corresponding author



(a) Public Fairness (b) Private Fairness (c) Partial Fairness with Belief-guided

Figure 1: An illustration for the fairness challenge in strategic classification. (a) Public fairness allows strategic behavior from advantaged agents, leading to fairness reversal *(left)*. (b) Private fairness avoids manipulation but rejects qualified individuals, causing social welfare loss *(center)*. (c) Our belief-guided strategic mechanism enables sequential learning for agents with partial fairness, improving group welfare and preventing fairness reversal *(right)*.

welfare, i.e., the acceptance rate of the qualified individuals, will decrease. Subsequently, reinforcing concealed fairness constraints leads to less incentivized model design, violating *the principle of incentive alignment* (Haghtalab et al. 2020). Combining the two sides, concurrent approaches on strategic fairness reveal a fundamental **dilemma**:

> *Fully public constraints can lead to fairness reversals and exacerbate inequality, while fully private constraints risk reducing overall social welfare.*

To address this dilemma, we introduce the novel problem *Partial Fairness Awareness* (PFA), a trade-off that balances the extremes of full transparency and privacy in fairness-aware strategic classification. In PFA, only a candidate set of fairness constraints is released to agents while the grounding constraint remains concealed. To fill this gap, we design a **belief-guided strategic mechanism**, where agents sequentially interact with the decision system and maintain a belief distribution over the candidate fairness constraints. At each round of interaction, the agent selects its optimal manipulation strategy based on its current belief, and updates the belief distribution with the feedback from the system (as shown in Figure 1(c)). As interactions proceed, this sequential learning process enables the agents to gradually infer and align with the actual (grounding) fairness constraint employed by the system. Our theory and empirical results demonstrate that this belief-guided mechanism effectively mitigates fairness reversal and improves social group welfare.

Our main contributions are summarized as follows:

- We conduct a systematic theoretical analysis of the fundamental dilemma between public and private disclosure of fairness constraints in strategic classification. To address the limitations of these two extremes, we formulate a novel problem of **Partial Fairness Awareness (PFA)**.
- To tackle the PFA problem, we design a **belief-guided strategic mechanism**, formally modeled via Bayesian inference. We rigorously prove that, under this mechanism, agents' beliefs converge to the true fairness constraint employed by the system, thereby improving both group fairness and overall social welfare.
- Extensive experiments on both synthetic and real-world datasets demonstrate that the PFA mechanism significantly outperforms conventional approaches based on fully public or fully private fairness constraints, with respect to fairness, social welfare, and predictive accuracy.

Related Work

Strategic Machine Learning

Strategic classification (Hardt et al. 2016) studies settings where individuals manipulate their features to influence model outcomes (Dong et al. 2017; Shavit, Edelman, and Axelrod 2020; Chen, Liu, and Podimata 2020; Harris, Heidari, and Wu 2021; Zrnic et al. 2021; Tsirtsis et al. 2024; Lv et al. 2025). Several works address challenges arising from unknown manipulations or limited agent information (Shao, Blum, and Montasser 2024; Ghalme et al. 2021; Yang et al. 2025a). Recent studies incorporate causal reasoning into strategic learning (Miller, Milli, and Hardt 2020; Chen, Wang, and Liu 2023; Horowitz and Rosenfeld 2023; Vo et al. 2024; Efthymiou et al. 2025; Chang et al. 2024; Yang et al. 2025b; Wang et al. 2022, 2023), distinguishing between manipulable and genuinely improvable features. This line of work emphasizes how strategic responses may reflect or distort true underlying qualifications. Relatedly, performative prediction (Perdomo et al. 2020; Hardt, Jagadeesan, and Mendler-Dünner 2022; Hardt and Mendler-Dünner 2023; Mendler-Dünner, Ding, and Wang 2022; Mofakhami, Mitliagkas, and Gidel 2023) analyzes how predictive models influence the data distribution over time through repeated deployment. A complementary direction focuses on promoting social welfare (Haghtalab et al. 2020; Estornell et al. 2023a; Xie and Zhang 2024), designing mechanisms that align agent incentives with collective benefit.

Fairness-aware Strategic Classification

Recent research has examined the complex interplay between fairness and strategic behavior in machine learning. Several studies highlight that unequal manipulation costs can exacerbate disparities even in the presence of fairness constraints (Hu, Immorlica, and Vaughan 2019; Milli et al. 2019). To mitigate these issues, various methods have been proposed, such as optimizing classifiers to reduce manipulation costs for disadvantaged groups (Keswani and Celis 2023; Wang et al. 2025) or employing minimax group fairness frameworks (Diana, Sharifi-Malvajerdi, and Vakilian 2024; Zheng et al. 2025). Other work evaluates fairness through agents' equilibrium behaviors (Shimao et al. 2025; Yang et al. 2024; Wang 2025) and explores how incentive

structures can influence manipulation (Zhang et al. 2022). Moreover, group fairness constraints may unintentionally result in fairness reversal when agents strategically modify their features (Estornell et al. 2023b). Recent efforts also focus on constructing fairness-aware models that anticipate manipulation and promote genuine improvement (Alhanouti and Naghizadeh 2025).

Preliminary

We present the background on strategic machine learning and social welfare. Throughout this paper, we denote random variables by uppercase letters (e.g., X and Y) and their realizations by lowercase letters (e.g., x and y). Bold symbols (e.g., $\mathbf{x}$ and $\mathbf{X}$) are used for vectors or matrices.

Strategic Classification

The strategic classification problem is modeled as a Stackelberg game (Li and Sethi 2017), where a **decision maker** defines a classification function $f : \mathbb{R}^d \to \{0, 1\}$, 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, Milli, and Hardt 2020). The optimal manipulated feature $\mathbf{x}'$ is determined by the best-response function $br(\mathbf{x})$:

Definition 1 (Strategic Manipulation)**.**

$$\mathbf{x}' = br(\mathbf{x}) = \arg\max_{x \in \mathcal{D}} U(\mathbf{x}, \mathbf{x}'), \tag{1}$$

where $\mathcal{D}$ denotes the distribution of the agents' features. Specifically, the utility function $U(\mathbf{x}, \mathbf{x}')$ is given by

$$U(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}') - \lambda c(\mathbf{x}, \mathbf{x}'), \tag{2}$$

where $f(\mathbf{x}') \in \{0, 1\}$ represents the classification outcome under the modified features and $c(\mathbf{x}, \mathbf{x}')$ is the cost associated with modification. And $\lambda > 0$ is a trade-off parameter that balances the classification benefit against the associated cost.

To mitigate strategic manipulation, the decision rule f' is optimized to maximize expected classification accuracy:

Definition 2 (Decision Optimization)**.**

$$f' \in \arg\max_{f \in \mathcal{F}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\mathbb{1}\big(f(br(\mathbf{x})) = y\big) \right], \tag{3}$$

where $\mathcal{F}$ is the set of feasible decision rules, $\mathbf{x}$ is the original feature vector in distribution $\mathcal{D}$, and y is the agent's label.

Fairness-aware Strategic Classification

In practical applications, strategic classification tasks are often required to satisfy group fairness constraints to ensure equitable outcomes for different demographic groups by imposing fairness constraints, such as demographic parity, on the classifier's decisions:

Definition 3 (Fairness-aware SC)**.** *Given fairness metric $\mathcal{C}$ (e.g., demographic parity gap) with group g, the strategic classifier f is optimized as:*

$$\begin{aligned} &\max_{f \in \mathcal{F}} \mathbb{E}_{(x, y, g) \sim D} \left[\mathbb{I}(f(br(\mathbf{x}; g)) = y) \right], \\ &\quad \textit{subject to} \quad \mathcal{C}(f; D, G) \leq \delta, \end{aligned} \tag{4}$$

where $\mathcal{C}(f; \mathcal{D}, G)$ quantifies the group fairness gap between groups in G (the set of all protected groups), and δ controls the allowable disparity. Moreover, $br(\mathbf{x}; g)$ denotes group-dependent strategic manipulation (best response) for agents in group g. This dependence on g arises because fairness constraints may affect groups differently, leading to distinct strategic manipulations.

Social Welfare in Strategic Classification

While traditional SC models emphasize robustness against manipulation, they often overlook the possibility that agents may engage in *genuine improvement*, that is, modifying their features in a way that reflects a true enhancement of their underlying qualification or ability (Horowitz and Rosenfeld 2023; Chen, Wang, and Liu 2023).

In many real-world settings, such improvements are desirable, e.g., a job applicant might complete a relevant training program to enhance employability. As a result, *genuinely improved* individuals may be misclassified and rejected, thereby discouraging effort and undermining long-term social welfare. More formally, we present the definition of *social welfare* to measure the number of truly qualified individuals who are accepted by the classifier after strategic manipulation (Haghtalab et al. 2020; Estornell et al. 2023a).

Definition 4 (Social Welfare)**.** *The social welfare W is defined as*

$$W = \sum_i \mathbb{I}(t_i \geq \tau_{qual}) \cdot \mathbb{I}(s(\mathbf{x}_i') \geq \theta_{g_i}), \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function, t_i is the true qualification score of agent i, τ_{qual} is the qualification threshold, $\mathbf{x}_i'$ is the strategically modified feature vector, $s(\cdot)$ is the classifier's scoring function that maps features to a decision score, and θ_{g_i} is the decision threshold for group g_i.

A higher value of W reflects better alignment between the classifier's decisions and agents' true qualifications, and thus greater societal benefit.

Dilemma of Strategic Fairness

Public Fairness: Reversal Phenomenon

On the one hand, common SC approaches always assume the public accessibility of the fairness constraints. However, existing theories already inform that *the advantaged group will utilize the exposure of fairness constraints to further manipulations*, as shown in Figure 2, leading to exacerbated inequality (i.e., compounded unfairness):

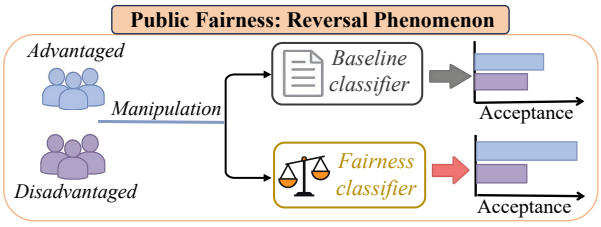


Figure 2: Illustration of the reversal phenomenon with public fairness in strategic classification.

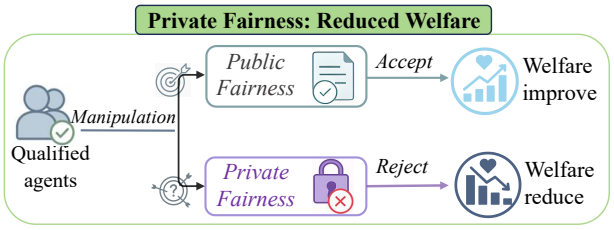


Figure 3: Illustration of reduced welfare: Private fairness constraints result in the rejection of qualified agents.

Lemma 1 (Fairness Reversal (Estornell et al. 2023b)). *Let $f^\dagger$ be a classifier trained to satisfy a group fairness constraint (e.g., demographic parity), and let $f^\sim$ be a baseline classifier without a fairness constraint. In the context of strategic manipulation, the reversal phenomenon occurs, defined as:*

$$\Delta_{fair}(f^\dagger, \mathcal{D}) > \Delta_{fair}(f^\sim, \mathcal{D}), \tag{6}$$

where $\Delta_{fair}(\cdot, \mathcal{D})$ denotes the fairness gap (e.g., difference in acceptance rates across groups) measured after strategic manipulation. That is, $f^\dagger$, though trained to be fair, exhibits a larger fairness gap than $f^\sim$ under agent manipulation.

Private Fairness: Reduced Welfare

On the other hand, a natural question falls in that *what if when the fairness constraints are fully private*, without any exposure to the agents. To address this, we further conduct an extensive theoretical analysis, informing that *concealing fairness constraints leads to a reduction in social welfare* (as defined in Eq. (5)), despite prior findings (Bent 2019; Mutlu, Yousefi, and Ozmen Garibay 2022) suggesting potential protection of group equality:

Theorem 1 (Welfare Reduction with Private Fairness [1]). *Let $\mathbb{E}[W_{public}]$ and $\mathbb{E}[W_{private}]$ denote the expected social welfare under public and private fairness constraint settings, respectively. Then:*

$$\mathbb{E}[W_{private}] < \mathbb{E}[W_{public}]. \tag{7}$$

Remark 1. Such welfare loss illustrates a practical drawback of private fairness settings, where even well-intentioned policies can unintentionally exclude qualified candidates and undermine the incentives for genuine improvement. As shown in Figure 3, the expected social welfare under private settings is always less than that under public settings.

Additional Manipulation Costs. Furthermore, when fairness criteria are private, agents struggle to accurately target the system's requirements, resulting in inefficient and often redundant manipulation efforts. This misalignment consequently leads to increased manipulation costs relative to the public setting.

Proposition 1 (Additional cost with private fairness [2]). *Hiding group-specific fairness constraints leads to higher manipulation costs for both advantaged and disadvantaged groups than disclosing them.*

[1]The detailed proof is included in Appendix A. The supplementary appendix is available in the arXiv version of this paper.

[2]The detailed proof is included in Appendix B.

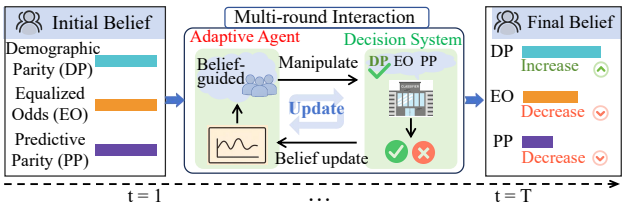


Figure 4: Illustration of belief-guided strategic mechanism: At each round, the agent updates its belief over the fairness mechanism set (e.g., *Demographic Parity, Equalized Odds*, and *Predictive Parity*) based on feedback from the decision system, enabling gradual alignment with the true fairness constraint.

Partial Fairness Awareness Formulation and Belief-Guided Mechanism

In this section, we formally define the *Partial Fairness Awareness* (PFA) problem in strategic classification, and present our belief-guided adaptation approach for learning and decision-making under this setting. We also analyze the impact of PFA and the belief-guided strategic mechanism on fairness-aware strategic classification.

Partial Fairness Awareness in Strategic Classification

To address the fairness dilemma in strategic classification, we naturally propose the viewpoint that *the fairness constraints should be partially exposed to the agents*, termed as ***P**artial **F**airness **A**wareness (PFA)*. To be specific, our PFA problem only lets agents access the set of possible fairness constraints (i.e., the information set), while leaving the fact that which constraint is indeed adopted by the decision-maker unobserved. Therefore, agents have to infer their own belief distributions from interactions with the decision-maker in the *sequential learning* process.

Definition 5 (Fairness Mechanism Set). *Let $\mathcal{M}$ denote the fairness mechanism set, i.e., the collection of all candidate fairness constraints the system may employ. The true, but unobserved, mechanism in effect is denoted by $\mathbf{m}^* \in \mathcal{M}$.*

For example, $\mathcal{M} = \{\texttt{None}, \texttt{DP}, \texttt{PP}, \texttt{EO}\}$ may represent the options of no constraint, demographic parity, and equalized odds, respectively, with $\mathbf{m}^* = \texttt{DP}$ indicating that demographic parity is adopted by the decision-maker.

Definition 6 (Partial Fairness Awareness). *Given a fairness mechanism set $\mathcal{M}$ comprising all plausible group fairness constraints the system may employ, and the true but unknown mechanism $\mathbf{m}^* \in \mathcal{M}$, the agent's knowledge is limited to the set $\mathcal{M}$, not the identity of $\mathbf{m}^*$.*

Consequently, based on the fairness mechanism set, we then define the concept of *belief distribution* over the fairness mechanism set, featuring the confidence of each agent in each candidate mechanism:

Definition 7 (Belief Distribution). *The belief distribution (probability distribution) is defined as $\mathbf{b} = \{b^{(m)}\}_{m \in \mathcal{M}}$, where each $b^{(m)}$ represents the agent's belief in the fairness mechanism $m \in \mathcal{M}$.*

Notably, our sequential setup allows multi-round interactions between agents and the decision-maker, with the time step denoted as $t \in [T]$. Hence, in the initial time step, the agents will have an initial belief in a set of possible mechanisms (e.g., $\mathbf{b}_0 = \{b_0^{(m)}\}_{m\in\mathcal{M}}$), and gradually refine this belief (e.g., $\mathbf{b}_t$) in each time step t through repeated interactions and observed decision results via the strategic process of *Belief-Guided Manipulation*:

Definition 8 (Belief-guided Manipulation)**.** *At each round t, the agent selects the belief-guided strategic manipulation with current belief $\mathbf{b}_t$, by maximizing expected utility under fairness uncertainty.*

This setup reflects our intuitions spanning over lots of realistic scenarios, where agents begin with no knowledge of the system's internal fairness mechanism and must learn over time through interaction (as shown in Figure 4).

Belief-Guided Strategic Mechanism for PFA

Within the concepts introduced above, we then detail the sequential policies of agents in each round, including *how the belief of fairness constraint* is updated, and *how agents strategically manipulate* based on the current belief.

Belief Initialization. At each round t, the agent maintains a normalized belief distribution $\mathbf{b}_t = \{b_t^{(m)}\}_{m\in\mathcal{M}}$, where $\sum_{m\in\mathcal{M}} b_t^{(m)} = 1$ and $b_t^{(m)} \in [0, 1]$. If no prior knowledge is assumed, the initial belief $\mathbf{b}_0$ is set uniformly over $\mathcal{M}$ (e.g., $b_{t=0}^{(m)} = 0.25$ for $|\mathcal{M}| = 4$).

Belief-guided Strategic Manipulation. At the round t, given the current belief $\mathbf{b}_t$, the agent aims to select a feature modification $\mathbf{x}' \in \mathcal{X}$ that maximizes their expected utility for getting favorable results. Formally, the agent solves the following optimization problem:

$$\mathbf{x}' = \arg\max_{\mathbf{x}'\in\mathcal{X}} \left(\mathbb{E}_{m\sim\mathbf{b}_t}\mathbb{P}_{f^{(m)}}(f^{(m)}(\mathbf{x}')=1) - \lambda\cdot c(\mathbf{x},\mathbf{x}')\right), \tag{8}$$

where $\mathbb{P}_{f^{(m)}}(f^{(m)}(\mathbf{x}') = 1)$ denotes the estimated probability of being accepted by the classifier $f^{(m)}$ under fairness mechanism $m \in \mathcal{M}$, and we use $f^{(m)}(\mathbf{x}') = 1$ because agents strategically optimize for acceptance rather than prediction accuracy. The trade-off parameter $\lambda > 0$ controls the agent's tolerance between acceptance reward and manipulation cost $c(\cdot)$.

Belief Update. After observing the decision feedback y_t for submitted input $\mathbf{x}'_t$, the agent updates its belief over the mechanism set $\mathcal{M}$. Specifically, we adopt a soft Bayesian update with exponentiated likelihoods:

$$\tilde{b}_{t+1}^{(m)} = \frac{b_t^{(m)} \cdot \left(\mathbb{P}_{f^{(m)}}(y_t \mid \mathbf{x}'_t)\right)^{\eta}}{\sum_{m'\in\mathcal{M}} b_t^{(m')} \cdot \left(\mathbb{P}_{f^{(m')}}(y_t \mid \mathbf{x}'_t)\right)^{\eta}}, \tag{9}$$

where $\eta > 0$ is a belief update rate controlling how strongly the agent reacts to new observations. A larger η results in faster concentration of belief, while a smaller η leads to more conservative updates.

The whole process of the belief-guided strategic mechanism is illustrated in Algorithm 1.

Algorithm 1: Belief-guided Strategic Mechanism

Require: Fairness mechanism set $\mathcal{M}$; total rounds T; belief update rate η
Initialize agent's belief $\mathbf{b}_0$ over $\mathcal{M}$
for $t = 0$ **to** T **do**
  Strategic Manipulation:
  Agent selects input $\mathbf{x}'_t$ by maximizing expected utility under current belief $\mathbf{b}_t$ with Eq. (8)
  Agent submits $\mathbf{x}'_t$ and receives decision feedback
  Belief Update:
  for each $m \in \mathcal{M}$ **do**
    Update temporary belief $\tilde{b}_{t+1}^{(m)}$ using Eq. (9)
  end for
end for
return Final belief distribution $\mathbf{b}_{t=T}$

Remark 2. The belief update step is formulated as a Bayesian update, where the likelihood $\mathbb{P}_{f^{(m)}}(y_t|\mathbf{x}'_t)$ represents the probability of observing the feedback y_t under each candidate mechanism m.

Theoretical Analysis

As the agent interacts with the decision system and updates beliefs, provided that the candidate mechanisms in $\mathcal{M}$ are statistically distinguishable, the agent's belief distribution $\mathbf{b}_t$ is guaranteed to converge to the true mechanism m^*. Therefore, we have the following theorem with the proof and convergence rate analysis provided in Appendix C.

Theorem 2 (Belief Convergence)**.** *Suppose the true mechanism is $m^* \in \mathcal{M}$, and each alternative $m \neq m^*$ is statistically distinguishable based on observed feedback. Then, for any $\epsilon > 0$, there exists $T > 0$ such that:*

$$b_t^{(m^*)} > 1 - \epsilon, \quad \textit{with } t > T. \tag{10}$$

Consequently, our PFA framework simultaneously addresses two key challenges in strategic classification: it prevents the immediate exploitation of group-specific fairness constraints (thus reducing fairness reversal), and it enables agents to gradually adapt via feedback, increasing the acceptance of truly qualified individuals and improving long-term social welfare. These advantages are formalized below.

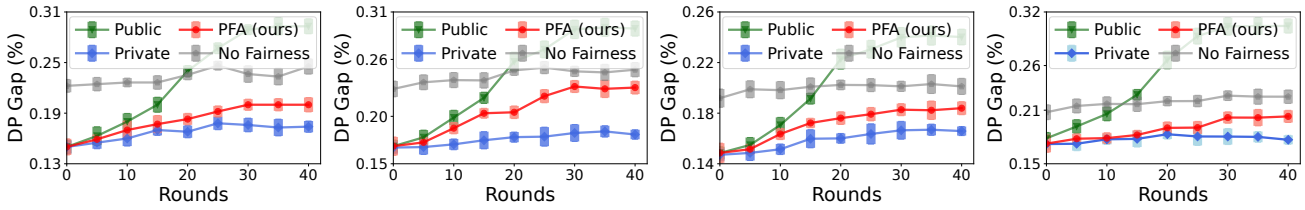
Theorem 3 (Fairness Gap Reduction under PFA [3])**.** *Given the fairness gap metric Δ_{fair} (e.g., the difference in true positive rates or acceptance rates between protected and advantaged groups). Let $\mathbb{E}[\Delta_{fair}^{PFA}]$ and $\mathbb{E}[\Delta_{fair}^{Public}]$ be the expected fairness gaps under the Partial Fairness Awareness (PFA) and fully public fairness mechanisms, respectively. Then,*

$$\mathbb{E}[\Delta_{fair}^{PFA}] < \mathbb{E}[\Delta_{fair}^{Public}]. \tag{11}$$

Theorem 4 (Improvement of Social Welfare under PFA)**.** *Let W_{PFA} and $W_{Private}$ denote the expected social welfare achieved under the PFA and private fairness settings, respectively, defined as the expected sum of qualified agents accepted minus the cost of manipulation. Then,*

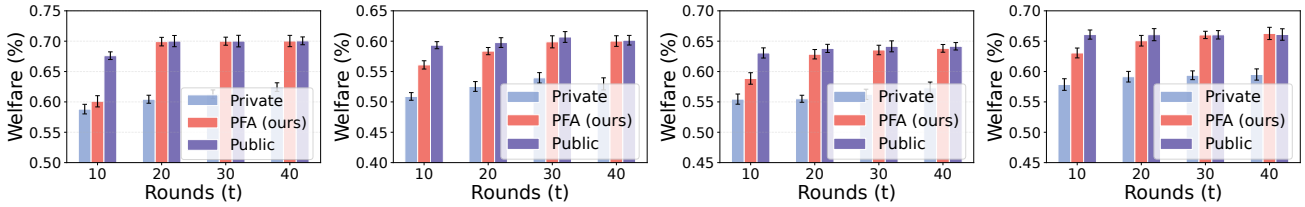
$$\mathbb{E}[W_{PFA}] > \mathbb{E}[W_{Private}], \tag{12}$$

[3]The proofs of Theorems are included in Appendices D and E.



(a) Adult Dataset (b) Credit Dataset (c) Diabetes Dataset (d) Synthetic Dataset

Figure 5: Performance of demographic parity gap on different real-world and synthetic datasets.



(a) Adult Dataset (b) Credit Dataset (c) Diabetes Dataset (d) Synthetic Dataset

Figure 6: Performance of social group welfare on different real-world and synthetic datasets.

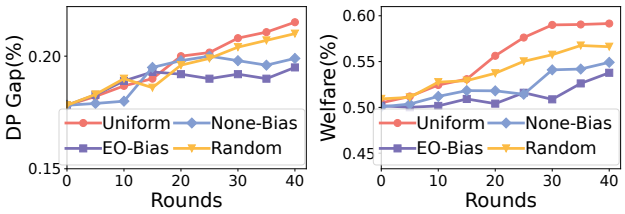
where the expectation is over agents' sequential learning and adaptation dynamics.

DP Gap$(\%)$	$t=5$	10	15	20	30	40
$\eta = 0.01$	0.170	0.172	0.175	0.182	0.186	0.188
$\eta = 0.05$	0.173	0.178	0.182	0.189	0.192	0.192
$\eta = 0.1$	0.181	0.183	0.186	0.190	0.193	0.193
$\eta = 0.5$	0.182	0.185	0.189	0.193	0.195	0.196

Table 1: Performance of demographic parity gap for different learning rates η with rounds T.

Welfare$(\%)$	$t=5$	10	15	20	30	40
$\eta = 0.01$	0.545	0.552	0.562	0.568	0.573	0.573
$\eta = 0.05$	0.550	0.561	0.568	0.573	0.580	0.581
$\eta = 0.1$	0.553	0.566	0.572	0.581	0.588	0.590
$\eta = 0.5$	0.561	0.578	0.585	0.589	0.596	0.597

Table 2: Performance of the group welfare (%) for different learning rates η with rounds T.



(a) DP Gap (b) Group Welfare

Figure 7: Ablation experimental results of the belief initialization distribution.

Experiment

Experimental Setup

Datasets. We evaluate our framework on four datasets, i.e., three real-world datasets and one synthetic benchmark:

- *Credit* (Yeh and Lien 2009): Credit card default prediction based on financial records, using gender as the sensitive attribute.
- *Adult* (Becker and Kohavi 1996): Income classification from census features; gender is used as the protected attribute.
- *Diabetes* (Teboul 2015): A medical dataset containing clinical and demographic attributes used to assess the risk of diabetes, using gender as the sensitive attribute.
- *Synthetic* (Lopez-Rojas, Elmir, and Axelsson 2016): Simulated mobile transaction data for fraud detection, with group membership as the sensitive attribute.

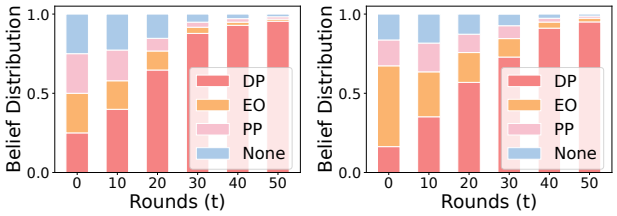
Metrics. We report fairness metrics and the group welfare metric. In fairness metrics, we mainly use the DP (demographic parity) Gap, but also employ EO (equalized odds) gap and PP (predictive parity) gap for more comprehensive evaluation.

- **DP Gap** measures the disparity in acceptance rates across groups:

$$\text{DP Gap} = |\mathbb{P}(\hat{y} = 1 \mid g = A) - \mathbb{P}(\hat{y} = 1 \mid g = B)|, \quad (13)$$

  where lower values indicate better statistical parity.
- **Group welfare**, defined as in Eq. (5), measuring the acceptance of truly qualified individuals across groups.

Fairness Settings. We compare three fairness strategies: (i)**public fairness**, where group-specific constraints are fully



(a) Uniform Initialization (b) EO-biased Initialization

Figure 8: Ablation experimental results of the belief initialization distribution.

disclosed to agents. (ii)**private fairness**, where all constraints are hidden and agents assume a single threshold. (iii) **Partial Fairness Awareness**, where agents gradually learn the underlying mechanism from decision feedback over multiple rounds.

Baseline. All mechanisms use the same linear classifier and Mahalanobis distance for manipulation cost (Gavish et al. 2021), ensuring only the transparency level differs. To evaluate alignment between classifier decisions and agent qualification, we use the *strategic improvement* framework (Miller, Milli, and Hardt 2020; Chen, Wang, and Liu 2023). The experimental results are presented in Figures 5 and 6.

Implementation Details

All experiments are implemented in Python 3.10 using the *scikit-learn* library for linear classifiers. The manipulation cost is computed as the Mahalanobis distance (Gavish et al. 2021), consistent across all settings. For PFA, belief updates use a multiplicative weights algorithm with learning rate $\eta = 0.002$. Each agent interacts with the classifier for up to $T = 40$ rounds. All experiments are conducted on a single NVIDIA TITAN V (12GB) GPU. More details and experimental results are provided in Appendix F.

Ablation Study

To investigate the effect of belief-guided learning in PFA, we conduct two ablation studies as follows. Unless otherwise stated, all other parameters are fixed as above.

Belief Update Rate. We vary the belief update rate η in $\{0.01, 0.05, 0.1, 0.5\}$ to evaluate its effect on convergence and adaptation stability. Beliefs are initialized uniformly over the candidate mechanism set $\mathcal{M}$, and agents interact for $T = 40$ rounds. Results are in Tables 1 and 2.

Initial Belief Distribution. To assess robustness to prior knowledge or bias, we experiment with non-uniform initial beliefs, including strong priors (e.g., EO-biased $b_{t=0}^{(\text{EO})} = 0.51$) and random weightings across $\mathcal{M}$. All other settings remain constant. Results are summarized in Figure 7.

Result Analysis

Main Results. Figure 5 summarizes the group fairness outcomes under different mechanisms, as measured by the Demographic Parity (DP) gap, across four representative datasets. Consistently, the **PFA (ours)** method achieves a markedly lower group-fairness gap compared to the public fairness. Notably, this improvement is robust across both real-world datasets (Adult, Credit, Diabetes) and the synthetic benchmark. As the number of interaction rounds increases, the DP gap for PFA remains stable and does not exhibit the sharp increases seen under Public or No Fairness settings, especially after $T > 20$, indicating strong resistance to fairness reversal. These findings provide empirical validation for Theorem 3.

Figure 6 further illustrates the corresponding dynamics of social group welfare under the same settings. Here, the **PFA (ours)** not only matches but often surpasses the welfare attained by the Public baseline, particularly in multiple rounds. In all datasets, social welfare improves steadily with additional rounds of interaction under PFA. This trend confirms that PFA enables agents to adapt their behavior in alignment with both fairness and qualification, effectively mitigating the typical welfare loss associated with private fairness constraints. These observations support Theorem 4.

As illustrated in Figure 8, we examine the evolution of agent beliefs under different initializations. In all cases, the belief distribution quickly converges toward the true fairness mechanism (DP) adopted by the jury, regardless of the initial setting. This empirical result aligns with Theorem 2, confirming that agents can reliably recover the underlying mechanism after a moderate number of interaction rounds, even when starting from a uniform or biased prior.

Ablation Results. The ablation studies evaluate the sensitivity of the PFA framework to two key factors: the belief update rate η and the initial belief distribution. As shown in Tables 1 and 2, intermediate learning rates ($\eta = 0.05$ or 0.1) achieve the best balance, yielding both low group fairness gaps and high social welfare. In contrast, very small learning rates slow down convergence, while overly large rates can introduce instability and degrade early performance.

Figure 7 further examines the effect of varying initial belief distributions. Despite some transient fluctuations in fairness metrics and welfare during early rounds, the belief consistently converges to the correct fairness mechanism and achieves comparable welfare across all settings. These findings demonstrate the robustness of the belief-guided mechanism to prior uncertainty in the initial belief distribution.

Overall, these results demonstrate that the PFA strikes a favorable balance between fairness and social welfare, maintaining equity across groups while supporting qualified individuals.

Conclusion

This work formulates the Partial Fairness Awareness (PFA) problem, addressing the fundamental trade-off between public and private fairness settings in strategic classification. To solve this problem, we develop a belief-guided strategic mechanism that enables agents to update their belief distribution of the fairness mechanism set based on feedback from the decision system. Theoretical analysis and experimental results on real-world and synthetic datasets demonstrate that PFA effectively mitigates fairness reversal and enhances social welfare in strategic machine learning. Future directions include extending PFA to multi-class settings and adapting to dynamic environments.

Ethics Statement

This work does not raise any ethical concerns. All experiments are conducted on publicly available datasets, and no human subjects or sensitive attributes are involved.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grants No. 62372459 and 62525213), the Natural Science Foundation of Heilongjiang Province (Grant No. LH2023C069), the Shanghai Sailing Program (Grant No. 24YF2711600), and the NUDT Youth Independent Innovation Science Fund (Grant No. ZK25-20).

References

Alhanouti, S.; and Naghizadeh, P. 2025. Anticipating Gaming to Incentivize Improvement: Guiding Agents in (Fair) Strategic Classification. *arXiv preprint arXiv:2505.05594*.

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository.

Bent, J. R. 2019. Is algorithmic affirmative action legal. *Geo. LJ*, 108: 803.

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

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

Chen, Y.; Wang, J.; and Liu, Y. 2023. Learning to Incentivize Improvements from Strategic Agents. *Transactions on Machine Learning Research*.

Diana, E.; Sharifi-Malvajerdi, S.; and Vakilian, A. 2024. Minimax Group Fairness in Strategic Classification. *arXiv preprint arXiv:2410.02513*.

Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4): 1–36.

Dong, J.; Roth, A.; Schutzman, Z.; Waggoner, B.; and Wu, Z. S. 2017. Strategic Classification from Revealed Preferences. arXiv:1710.07887.

Efthymiou, V.; Podimata, C.; Sen, D.; and Ziani, J. 2025. Incentivizing Desirable Effort Profiles in Strategic Classification: The Role of Causality and Uncertainty. *arXiv preprint arXiv:2502.06749*.

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

Estornell, A.; Das, S.; Liu, Y.; and Vorobeychik, Y. 2023b. Group-Fair Classification with Strategic Agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 389–399. New York, NY, USA: Association for Computing Machinery.

Gavish, M.; Talmon, R.; Su, P.-C.; and Wu, H.-T. 2021. Optimal Recovery of Precision Matrix for Mahalanobis Distance from High Dimensional Noisy Observations in Manifold Learning. arXiv:1904.09204.

Ghalme, G.; Nair, V.; Eilat, I.; Talgam-Cohen, I.; and Rosenfeld, N. 2021. Strategic Classification in the Dark. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 3672–3681.

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

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

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

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

Harris, K.; Heidari, H.; and Wu, S. Z. 2021. 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, 28728–28741. Curran Associates, Inc.

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

Hu, L.; Immorlica, N.; and Vaughan, J. W. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268.

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

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

Kučak, D.; Juričić, V.; and DJambić, G. 2018. Machine Learning in Education: A Survey of Current Research Trends. *Annals of DAAAM & Proceedings*, 29: 0441–0446.

Li, T.; and Sethi, S. P. 2017. A Review of Dynamic Stackelberg Game Models. *Discrete & Continuous Dynamical Systems-Series B*, 22(1).

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

Lv, X.; Mao, Y.; Li, H.; Liang, K.; Yang, J.; Huang, W.; Chi, H.; Chen, H.; Lan, L.; Yang, W.; and Wang, H. 2025. Breaking the Gradient Barrier: Unveiling Large Language Models for Strategic Classification. In *Advances in Neural Information Processing Systems*.

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

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

Milli, S.; Miller, J.; Dragan, A. D.; and Hardt, M. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 230–239.

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

Mutlu, E. Ç.; Yousefi, N.; and Ozmen Garibay, O. 2022. Contrastive counterfactual fairness in algorithmic decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 499–507.

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

Roemer, J. E.; and Trannoy, A. 2015. Equality of opportunity. In *Handbook of income distribution*, volume 2, 217–300. Elsevier.

Sánchez-Monedero, J.; Dencik, L.; and Edwards, L. 2020. 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*, 458–468.

Shao, H.; Blum, A.; and Montasser, O. 2024. Strategic classification under unknown personalized manipulation. *Advances in Neural Information Processing Systems*, 36.

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

Shimao, H.; Khern-Am-Nuai, W.; Kannan, K.; and Cohen, M. C. 2025. Strategic Best-Response Fairness Framework for Fair Machine Learning. *Information Systems Research*.

Singh, M. K.; and Kulkarni, A. A. 2024. Optimal Stochastic Decision Rule for Strategic Classification. In *2024 National Conference on Communications (NCC)*, 1–6. IEEE.

Strathern, M. 1997. 'Improving ratings': audit in the British University system. *European Review*, 5(3): 305–321.

Teboul, A. 2015. Diabetes Health Indicators Dataset.

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

Vo, K. Q.; Aadil, M.; Chau, S. L.; and Muandet, K. 2024. Causal Strategic Learning with Competitive Selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15411–15419.

Wang, H.; Kuang, K.; Chi, H.; Yang, L.; Geng, M.; Huang, W.; and Yang, W. 2023. Treatment effect estimation with adjustment feature selection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2290–2301.

Wang, H.; Li, H.; Zou, H.; Chi, H.; Lan, L.; Huang, W.; and Yang, W. 2025. Effective and Efficient Time-Varying Counterfactual Prediction with State-Space Models. In *The Thirteenth International Conference on Learning Representations*.

Wang, H.; Yang, W.; Yang, L.; Wu, A.; Xu, L.; Ren, J.; Wu, F.; and Kuang, K. 2022. Estimating individualized causal effect with confounded instruments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1857–1867.

Wang, M.-z. 2025. SimProF: A Simple Probabilistic Framework for Unsupervised Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21153–21161.

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

Yang, S.; Yu, W.; Yang, W.; Liu, X.; Tan, H.; Lan, L.; and Xiao, N. 2025a. WildVideo: Benchmarking LMMs for Understanding Video-Language Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, W.; Lv, X.; Mao, Y.; Xu, L.; Jin, R.; Chen, H.; Ren, J.; Yang, J.; Chen, Y.; and Wang, H. 2025b. Advanced Strategic Improvement with Decision Interactions. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, 426–444. Cham: Springer Nature Switzerland.

Yang, W.; Wang, H.; Li, H.; Zou, H.; Jin, R.; Kuang, K.; and Cui, P. 2024. Your neighbor matters: Towards fair decisions under networked interference. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3829–3840.

Yeh, I.-C.; and Lien, C.-h. 2009. The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Syst. Appl.*, 36(2): 2473–2480.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *International Conference on Machine Learning*, 325–333.

Zhang, X.; Khalili, M. M.; Jin, K.; Naghizadeh, P.; and Liu, M. 2022. Fairness interventions as (dis) incentives for strategic manipulation. In *International Conference on Machine Learning*, 26239–26264. PMLR.

Zheng, C.; Pan, H.; Zhang, Y.; and Li, H. 2025. Adaptive Structure Learning with Partial Parameter Sharing for Post-Click Conversion Rate Prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, 233–243. Association for Computing Machinery.

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