Non-linear Welfare-Aware Strategic Learning

Tian Xie[1], Xueru Zhang[1]

[1]Department of Computer Science and Engineering, the Ohio State University
{xie.1379, zhang.12807}@osu.edu

Abstract

This paper studies algorithmic decision-making in the presence of strategic individual behaviors, where an ML model is used to make decisions about human agents and the latter can adapt their behavior strategically to improve their future data. Existing results on strategic learning have largely focused on the *linear* setting where agents with linear labeling functions best respond to a (noisy) linear decision policy. Instead, this work focuses on general non-linear settings where agents respond to the decision policy with only "local information" of the policy. Moreover, we simultaneously consider objectives of maximizing *decision-maker welfare* (model prediction accuracy), *social welfare* (agent improvement caused by strategic behaviors), and *agent welfare* (the extent that ML underestimates the agents). We first generalize the agent best response model in previous works to the non-linear setting and then investigate the compatibility of welfare objectives. We show the three welfare can attain the optimum simultaneously only under restrictive conditions which are challenging to achieve in non-linear settings. The theoretical results imply that existing works solely maximizing the welfare of a subset of parties usually diminish the welfare of others. We thus claim the necessity of balancing the welfare of each party in non-linear settings and propose an *irreducible* optimization algorithm suitable for general strategic learning. Experiments on synthetic and real data validate the proposed algorithm.

1 Introduction

When machine learning (ML) models are used to automate decisions about human agents (e.g., in lending, hiring, and college admissions), they often have the power to affect agent behavior and reshape the population distribution. With the (partial) information of the decision policy, agents can change their data strategically to receive favorable outcomes, e.g., they may exert efforts to genuinely improve their labels or game the ML models by manipulating their features without changing the labels. This phenomenon has motivated an emerging line of work on *Strategic Classification* (Hardt et al. 2016), which explicitly takes the agents' strategic behaviors into account when learning ML models.

One challenge in strategic classification is to model the interactions between agents and the ML decision-maker, as well as understand the impacts they each have on the other.

As the inputs of the decision system, agents inevitably affect the utility of the decision-maker. Meanwhile, the deployed decisions can further influence agents' future data and induce societal impacts. Thus, to facilitate socially responsible algorithmic decision-making, we need to consider the welfare of all parties when learning the decision system, including the welfare of the decision-maker, agent, and society. Specifically, the decision-maker typically cares more about assigning correct labels to agents when making decisions, while society will be better off if the decision policy can incentivize agents to improve their actual states (labels). For the agents, it is unrealistic to assign each of them a positive decision, but it is reasonable for agents to expect a "just" decision that does not underestimate their actual states.

However, most existing works on strategic classification solely consider the welfare of one party. For example, Dong et al. (2018); Braverman and Garg (2020); Jagadeesan, Mendler-Dünner, and Hardt (2021); Sundaram et al. (2021); Zhang et al. (2022); Eilat et al. (2022) only focus on *decision-maker welfare* by designing accurate classifiers robust to strategic manipulation. Kleinberg and Raghavan (2020); Harris, Heidari, and Wu (2021); Ahmadi et al. (2022a) solely consider maximizing *social welfare* by designing an optimal policy to incentivize the maximum improvement of agents. To the best of our knowledge, only a few works consider both *decision-maker welfare* and *social welfare* (Bechavod et al. 2022; Barsotti, Kocer, and Santos 2022). Among them, Bechavod et al. (2022) initiate a theoretical study to understand the relationship between the two welfare, where they identify a sufficient condition under which the *decision-maker welfare* and *social welfare* can be optimized at the same time. This paper still focuses on incentivizing agent improvement to maximize *social welfare*.

Moreover, most prior works limit the scope of analysis to linear settings by assuming both the agent labeling function and the ML decision rules are linear (Dong et al. 2018; Chen, Liu, and Podimata 2020; Ahmadi et al. 2021; Horowitz and Rosenfeld 2023). This is unrealistic in real applications when the relationship between features and labels is highly intricate and non-linear. Meanwhile, previous works typically modeled strategic behaviors as "(noisy) best responses" to the decision policy (Hardt et al. 2016; Jagadeesan, Mendler-Dünner, and Hardt 2021; Dong et al. 2018; Chen, Liu, and Podimata 2020; Ahmadi et al. 2021;

Horowitz and Rosenfeld 2023), which is not practical when the decision policy is highly complex and agents' knowledge about the policy (e.g., a neural network) is far from complete. It turns out that the existing theoretical results and methods proposed under linear settings where agents "best respond" can easily be violated in non-linear settings. Thus, there remain gaps and challenges in developing a comprehensive framework for strategic learning under general non-linear settings.

To tackle the limitations in prior works, we study the welfare of all parties in strategic learning under general non-linear settings where both the ground-truth labeling function and the decision policies are complex and possibly non-linear. Importantly, we develop a generalized agent response model that allows agents to respond to policy only based on local information/knowledge they have about the policy. Under the proposed response model, we simultaneously consider: 1) *decision-maker welfare* that measures the prediction accuracy of ML policy; 2) *social welfare* that indicates the level of agent improvement; 3) *agent welfare* that quantifies the extent to which the policy under-estimates the agent qualifications.

Under general non-linear settings and the proposed agent response model, we first revisit the theoretical results in Bechavod et al. (2022) and show that those results can easily be violated under more complex settings. We then explore the relationships between the objectives of maximizing decision-maker, social, and agent welfare. The results show that the welfare of different parties can only be aligned under rigorous conditions depending on the ground-truth labeling function, agents' knowledge about the decision policy, and the realizability (i.e., whether the hypothesis class of the decision policy includes the ground-truth labeling function). It implies that existing works that solely maximize the welfare of a subset of parties may inevitably diminish the welfare of the others. We thus propose an algorithm to balance the welfare of all parties.

The rest of the paper is organized as follows. After reviewing the related work in Sec. 2, we formulate a general strategic learning setting in Sec. 3, where the agent response and the welfare of the decision-maker, agent, and society are presented. In Sec. 4, we discuss how the existing theoretical results derived under linear settings are violated under non-linear settings and provide a comprehensive analysis of the decision-maker/agent/social welfare; we show all the conditions when they are guaranteed to be compatible in Thm. 4.1. This demonstrates the necessity of an "irreducible" optimization framework balancing the welfare of all parties, which we introduce in Sec. 5. Finally, in Sec. 6, we provide comprehensive experimental results to compare the proposed method with other benchmark algorithms on both synthetic and real datasets.

2 Related Work

Next, we briefly review the related work and discuss the differences with this paper.

2.1 Strategic Learning

Hardt et al. (2016) was the first to formulate strategic classification as a Stackelberg game where the decision-maker publishes a policy first following which agents best respond. A large line of subsequent research has focused on designing accurate classifiers robust to strategic behaviors. For example, Ben-Porat and Tennenholtz (2017); Chen, Wang, and Liu (2020); Tang, Ho, and Liu (2021) designed robust linear regression predictors for strategic classification, while Dong et al. (2018); Ahmadi et al. (2021); Lechner, Urner, and Ben-David (2023); Shao, Blum, and Montasser (2024) studied online learning algorithms for the problem. Levanon and Rosenfeld (2021) designed a protocol for the decision maker to learn agent best response when the decision policy is non-linear, but it assumes agents are perfectly aware of the complex classifiers deployed on them. Braverman and Garg (2020); Jagadeesan, Mendler-Dünner, and Hardt (2021) relaxed the assumption that the best responses of agents are deterministic and studied how the decision policy is affected by randomness. Sundaram et al. (2021) provided theoretical results on PAC learning under the strategic setting, while Levanon and Rosenfeld (2022); Horowitz and Rosenfeld (2023) proposed a new loss function and algorithms for linear strategic classification. Miller, Milli, and Hardt (2020); Shavit, Edelman, and Axelrod (2020); Alon et al. (2020) focused on causal strategic learning where features are divided into "causal" and "non-causal" ones. However, all these works study strategic classification under linear settings and the primary goal is to maximize *decision-maker welfare*. In contrast, our work considers the welfare of not only the decision-maker but also the agent and society. Another line of work related to strategic learning is *Performative Prediction* (Perdomo et al. 2020; Jin et al. 2024) to capture general model-dependent distribution shifts. Despite the generality, the framework is highly abstract and hard to interpret when considering the welfare of different parties. We refer to Hardt and Mendler-Dünner (2023) as a comprehensive survey.

2.2 Promote Social Welfare in Strategic Learning

Although the definition of social welfare is sometimes subtle (Florio 2014), *social welfare* is often regarded as the improvement of ground truth qualifications of agents in strategic learning settings and we need to consider how decision rules impact *social welfare* by shaping agent best response. A rich line of works are proposed to design mechanisms to incentivize agents to improve (e.g., Liu et al. (2019); Shavit, Edelman, and Axelrod (2020); Alon et al. (2020); Zhang et al. (2020); Chen, Wang, and Liu (2020); Kleinberg and Raghavan (2020); Ahmadi et al. (2022a,b); Jin et al. (2022); Xie and Zhang (2024)). The first prominent work was done by Kleinberg and Raghavan (2020), where they studied the conditions under which a linear mechanism exists to incentivize agents to improve their features instead of manipulating them. Specifically, they assume each agent has an effort budget and can choose to invest any amount of effort on either improvement or manipulation, and the mechanism aims to incentivize agents to invest all available efforts in improvement. Harris, Heidari, and Wu (2021) designed a

more advanced stateful mechanism considering the accumulative effects of improvement behaviors. They stated that improvement can accumulate over time and benefit agents in the long term, while manipulation cannot accumulate state to state. On the contrary, Estornell, Das, and Vorobeychik (2021) considered using an audit mechanism to disincentivize manipulation which instead incentivize improvement. Besides mechanism design, Rosenfeld et al. (2020) studied incentivizing improvement in machine learning algorithms by adding a look-ahead regularization term to favor improvement. Chen, Wang, and Liu (2020) divided the features into immutable, improvable, and manipulable features and explored linear classifiers that can prevent manipulation and encourage improvement. Jin et al. (2022) also focused on incentivizing improvement and proposed a subsidy mechanism to induce improvement actions and improve social well-being metrics. Bechavod et al. (2021) demonstrated the ability of strategic decision-makers to distinguish features influencing the label of individuals under an online setting. Ahmadi et al. (2022a) proposed a linear model where strategic manipulation and improvement are both present. Barsotti, Kocer, and Santos (2022) conducted several empirical experiments when improvement and manipulation are possible and both actions incur a linear deterministic cost.

The works closest to ours are Bechavod et al. (2022); Rosenfeld et al. (2020). Bechavod et al. (2022) provided analytical results of maximizing total improvement without harming anyone's current qualification, but they assumed both the ground truth and the decision rule are linear. Rosenfeld et al. (2020) maximized accuracy while punishing the classifier only if it brings the negative externality in nonlinear settings. However, solely punishing negative externality is far from ideal. Our work goes further by analyzing the relationships between the welfare of different parties and providing an optimization algorithm to balance the welfare.

2.3 Welfare-Aware Machine Learning

The early work on welfare-aware machine learning typically considers balancing *social welfare* and another competing objective such as fairness or profit. For example, Ensign et al. (2018) focused on long-term *social welfare* when myopically maximizing efficiency in criminal inspection practice creates a runaway feedback loop. Liu et al. (2020); Zhang et al. (2020); Guldogan et al. (2022) showed that instant fairness constraints may harm *social welfare* in the long run since the agent population may react to the classifier unexpectedly. Rolf et al. (2020) proposed a framework to analyze the Pareto-optimal policy balancing the prediction profit and a possibly competing welfare metric but assumed a simplified setting where the welfare of admitting any agent is constant which is not the case in strategic learning.

3 Problem Formulation

Consider a population of strategic agents with features $X = (X_1, ..., X_d)$, $\text{dom}(X) \subset \mathbb{R}^d$, and label $Y \in \{0, 1\}$ that indicates agent qualification for certain tasks (with "0" being unqualified and "1" qualified). Denote P_X as the probability density function of X, and assume labeling func-

tion $h(x) := P(Y = 1|X = x)$ is continuous. The agents are subject to certain algorithmic decisions $D(x) : \mathbb{R}^d \rightarrow \{0, 1\}$ (with "1" being positive and "0" negative), which are made by a decision-maker using a scoring function $f(x) := P(D(x) = 1|X = x)$. Unlike many studies in strategic classification that assume $f \in \mathcal{F}$ is a linear function[1], we allow $\mathcal{F}$ to be non-linear. Moreover, it is possible that hypothesis class $\mathcal{F}$ does not include h.

3.1 Information level and generalized agent response

We consider practical settings where the decisions deployed on agents have downstream effects and can reshape the agent's future data. We also focus on "benign" agents whose future labels change according to the fixed labeling function $h(x)$ (Raab and Liu 2021; Rosenfeld et al. 2020; Guldogan et al. 2022). Given a function f, we define agent best response to f in Def. 3.1, which is consistent with previous works on strategic learning (Hardt et al. 2016; Dong et al. 2018; Ahmadi et al. 2021).

Definition 3.1 (Agent best response to a function). Assume an agent with feature x will incur cost $c(x, x')$ by moving her feature from x to x'. Then we say the agent best responds to function f if she changes her features from x to $x^* = \arg\max_{x'}\{f(x') - c(x, x')\}$.

It is worth noting that most existing works in strategic learning assume agents either best respond to the exact f (e.g., Hardt et al. (2016); Horowitz and Rosenfeld (2023); Dong et al. (2018); Ahmadi et al. (2021)) or noisy f with additive noise (e.g., Levanon and Rosenfeld (2022); Jagadeesan, Mendler-Dünner, and Hardt (2021)). However, the best response model in Def. 3.1 requires the agents to know the exact values of $f(x')$ for all x' in the feature domain, which is a rather strong requirement and unlikely holds in practice, especially when policy $f \in \mathcal{F}$ is complex (e.g., logistic regression/neural networks). Instead, it may be more reasonable to assume agents only have "local information" about the policy f, i.e., agents with features x only have some information about f at x and they know how to improve their qualifications based on their current standings (Zhang et al. 2022; Raab and Liu 2021). To formalize this, we define the agent's *information level* and the resulting agent response in Def. 3.2.

Definition 3.2 (Information level and agent response). Agents with features x and *information level* K know the gradient of f at $x = [x_1, ..., x_d]$ up to the Kth order (the agents know $\{\nabla f(x), \nabla^2 f(x), ..., \nabla^K f(x)\}$ if all these gradients exist). Let Q_x^f be the Taylor expansion of f at x up to the Kth order. When f is only $k < K$ times differentiable, Q_x^f simply uses all existing gradient information. Then the agents with information level K will respond to f by first estimating f using Q_x^f, and then best responding to Q_x^f.

Def. 3.2 specifies the *information level* of agents and how they respond to the decision policy using all the information

---

[1] A linear scoring function can be $\mathbf{1}(w^T x + b \geq 0)$ (e.g., Levanon and Rosenfeld (2022); Horowitz and Rosenfeld (2023)) or $w^T x + b$ when $\text{dom}(X)$ is bounded (e.g., Bechavod et al. (2022)).

available at the locality. A lower level of K is more interpretable in practice. For example, agents with information level $K = 1$ use the first-order gradient of f to calculate a linear approximation, which means they only know the most effective direction to improve their current feature values. In contrast, agents with information level $K = 2$ also know how this effective direction to improve will change (i.e., the second-order gradient that reveals the curvature of f).

It is worth noting that the agent response under $K = 1$ is also similar to Rosenfeld et al. (2020) where agents respond to the decision policy by moving toward the direction of the local first-order gradient, and Dean et al. (2023) where the agents aim to reduce their risk based on the available information at the current state. Moreover, Def. 3.2 can also be regarded as a generalization of the previous models where agents best respond to the exact decision policy under the linear setting. To illustrate this, assume the agents have $K \geq 1$ and the linear policy is $f(x) = \beta^T x$, then the agents with feature x will estimate f using $Q_x^f(x') = f(x) + \nabla f(x) \cdot (x' - x) = \beta^T x'$, which is the same as f. However, for complex f, agents with information level K may no longer best respond to f. Prop. 3.3 illustrates the necessary and sufficient conditions for an agent to best respond to f.

Proposition 3.3. *An agent with information level K can best respond to f if and only if f is a polynomial with an order smaller than or equal to K.*

Prop. 3.3 shows that agents with only local information of policy can only best respond to f that is polynomial in feature vector X. In practical applications, f often does not belong to the polynomial function class (e.g., a neural network with *Sigmoid* activation may have infinite order gradients, and the one with *Relu* activation may be a piece-wise function), making it unlikely for agents to best respond to f. Note that although Def. 3.2 specifies the agents to best respond to Q_x^f which is an approximation of actual policy f using the "local information" at x, we also present a protocol in Alg. 2 to learn a more general response function $x^* = \Delta_\phi(x, I_k(f, x))$ from population dynamics where ϕ are learnable parameters and $I_k(f, x)$ is all information available for agents with information level K and feature x. This protocol can learn a rich set of responses when the agents have information level K and use their information to respond to f strategically.

3.2 Welfare of different parties

In this paper, we jointly consider the welfare of the decision-maker, agents, and society. Before exploring their relations, we first introduce the definition of each below.

Decision-maker welfare. It is defined as the utility the decision maker immediately receives from the deployed decisions, without considering the downstream effects and agent response. Formally, for agents with data (x, y), we define the utility the decision-maker receives from making the decision $D(x)$ as $u(D(x), y)$. Then the *decision-maker welfare* is just the expected utility over the population, i.e.,

$$\text{DW}(f) = \mathbb{E}_{x,y \sim P_{XY}}[u(D(x), y)].$$

where P_{XY} is the joint distribution of (X, Y). When $u(D(x), y) = \mathbf{1}(D(x) = y)$, $\text{DW}(f)$ reduces to the prediction accuracy. In Sec. 5 and 6, we focus on this special case, but the results and analysis can be extended to other utility functions.

Social welfare. It evaluates the impact of the agent response on population qualifications. We consider two types of social welfare commonly studied in strategic learning: *agent improvement* and *agent safety*. Specifically, agent improvement is defined as:

$$\text{IMP}(f) = \mathbb{E}_{x \sim P_X}[h(x^*) - h(x)].$$

$\text{IMP}(f)$ measures the increase in the overall population qualification caused by the agent response. This notion has been theoretically and empirically studied under linear settings (e.g., Kleinberg and Raghavan (2020); Bechavod et al. (2022)).

For agent safety, we define it as:

$$\text{SF}(f) = \mathbb{E}_{x \sim P_X}\left[\min\left\{h(x^*) - h(x), 0\right\}\right].$$

It illustrates whether there are agents suffering from deteriorating qualifications after they best respond. This notion has been studied in Bechavod et al. (2022) and was interpreted as "do no harm" constraint, i.e., the deployed decisions should avoid adversely impacting agents. Rosenfeld et al. (2020) has shown that *agent safety* can not be automatically guaranteed even under linear settings, and they proposed an algorithm to ensure safety in general settings. However, they did not consider agent improvement which we believe is an equally important measure of social welfare.

Agent welfare. Previous literature rarely discussed the welfare of agents. The closest notion was *social burden* (Milli et al. 2019) defined as the minimum cost a positively-labeled user must pay to obtain a positive decision, where the authors assume agents can only "game" the decision-maker without improving their true qualifications. Since the agents can always strive to improve in our setting, their efforts are not necessarily a "burden". However, a relatively accurate policy may intentionally underestimate the current qualification to incentivize agent improvement. Thus, we propose *agent welfare* which measures the extent to which the agent qualification is under-estimated by the decision-maker, i.e.,

$$\text{AW}(f) = \mathbb{E}_{x \sim P_X}\left[\min\left\{f(x) - h(x), 0\right\}\right].$$

When the policy underestimates the agent's true potential ($f(x) < h(x)$), the agent may be treated unfairly by the deployed decision (i.e., when a qualified agent gets a negative decision). We thus use $\text{AW}(f)$ to measure agent welfare.

4 Welfare Analysis

This section explores the relationships between the decision-maker, social, and agent welfare introduced in Sec. 3. Among all prior works, (Bechavod et al. 2022) provides a theoretical analysis of *social* and *decision-maker welfare* under linear settings. In this section, we first revisit the results in (Bechavod et al. 2022) and show that they may not be applicable to non-linear settings. Then, we examine the relationships between three types of welfare generally.

4.1 Beyond the linear setting

Bechavod et al. (2022) considered linear settings when both the labeling function h and decision policy f are linear and revealed that *decision-maker welfare* ($\mathrm{DW}(f)$) and *social welfare* ($\mathrm{IMP}(f)$) can attain the optimum simultaneously when the decision maker deploys $f = h$ and agents best respond to f. However, under general non-linear settings, the compatibility of welfare is rarely achieved and we present the necessary and sufficient conditions to guarantee the alignment of all welfare in Thm. 4.1.

Theorem 4.1 (Alignment of all welfare). *Suppose agents have information level K, then decision-maker welfare, social welfare, and agent welfare can attain maximum simultaneously regardless of the feature distribution P_X and cost function $c(x, x')$ if and only if h is polynomial in X with order smaller than or equal to K and $h \in \mathcal{F}$.*

Thm. 4.1 can be proved based on Prop. 3.3. Intuitively, *decision-maker welfare* and *agent welfare* are guaranteed to be maximized when f is accurate, while agents are only guaranteed to gain the largest improvement under an accurate f if they have a perfect understanding of f. When Thm. 4.1 is not satisfied, the welfare of different parties is incompatible under the non-linear settings; we use two examples to illustrate this below.
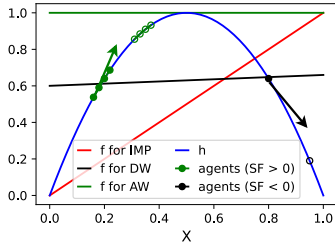

Figure 1: Illustration of the optimal linear policies and agent response when h (the blue curve) is quadratic and f is linear. There are four agents with features below 0.5 (green solid circle) and one agent with the feature above 0.5 (black solid circle). We observe that (i) the red line maximizes agent improvement $\mathrm{IMP}(f)$ but it is **unsafe** by only encouraging the green agents to improve while the black agent's qualification deteriorates. The green/black arrows and the hollow circles show how agent response improves/worsens their qualifications; (ii) the black line aims to maximize $\mathrm{DW}(f)$ by running a least square regression with respect to ground truth h, which is different from the red line; (iii) Neither the red nor the black line maximizes the *agent welfare* $\mathrm{AW}(f)$; only predictors above $h(x)$ (e.g., constant predictor $f(x) = 1$ shown by green line) achieves the optimal $\mathrm{AW}(f)$.

Example 4.2 (Quadratic labeling function with linear policy). *Consider a hiring scenario where a company needs to assess each applicant's fitness for a position. Each applicant has a feature $X \in [0, 1]$ demonstrating his/her ability (e.g., coding ability for a junior software engineer position). Suppose the labeling function $h(x) = -4x \cdot (x - 1)$*

is quadratic with maximum taken at $x = 0.5$, which implies that an applicant's fitness will decrease if the ability is too low ("underqualified") or too high ("overqualified")[2]. Under the quadratic labeling function, if the decision-maker can only select a decision policy f from a linear hypothesis class $\mathcal{F}$ (i.e., violating the requirement that $h \in \mathcal{F}$ in Thm. 4.1), then the optimal policies maximizing decision-maker welfare and agent improvement are completely different, i.e., $\mathrm{DW}(f)$ and $\mathrm{IMP}(f)$ cannot be maximized simultaneously. Moreover, agent safety $\mathrm{SF}(f)$ is also not guaranteed. To illustrate this, consider the example in Fig. 1 with five agents: four green agents with features below 0.5 and a black one above 0.5, then the policy that maximizes $\mathrm{IMP}(f)$ can cause the qualification of black agent to deteriorate and lead to negative agent safety $\mathrm{SF}(f) < 0$.

Example 4.3 (Quadratic labeling function and policy). *Under the same setting of Example 4.2, instead of restricting f to be linear, let the decision-maker deploy $f(x) = h(x) = -4x \cdot (x - 1)$. Obviously, f maximizes both the decision-maker welfare and agent welfare. However, it can hurt social welfare. Consider an example shown in Fig. 2 where all agents have the same feature $x = 0.4$ and information level $K = 1$ (violating the information level requirement of Thm. 4.1). Their cost function is $c(x, x') = (x' - x)^2$. Since $\nabla f(x) = -8x + 4 = 0.8$, agents derive $Q_x^f(x') = f(x) + 0.8 \cdot (x' - x) = 0.8x' + 0.64$. Then agent post-response feature will be $x^* = x + \frac{0.8}{2} = 0.8$. As a result, the agent qualification becomes $h(x^*) = h(0.8) = 0.64 < h(0.4) = 0.96$, which is worse than the initial qualification, i.e., agent safety is not optimal by deploying $f = h$.*
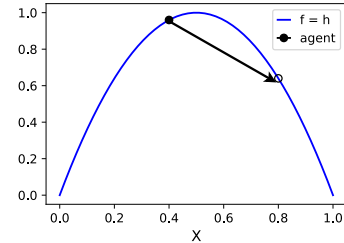

Figure 2: Violation of agent safety $\mathrm{SF}(f)$ when $f = h$ is quadratic (the blue curve): All agents have the same feature value concentrating on $X = 0.4$ (the solid black circle). If applying the ground truth as the decision policy, all agents will respond by moving to $x^* = 0.8$ (the hollow black circle). Although the policy maximizes $\mathrm{AW}(f)$ and $\mathrm{DW}(f)$, all agents become less qualified after they respond, harming the *social welfare*. Both $\mathrm{IMP}(f)$ and $\mathrm{SF}(f)$ are smaller than 0.

Examples 4.2 and 4.3 illustrate the potential conflict between decision-maker, social, and agent welfare and verify Thm. 4.1. Even for simple quadratic settings, the results under the linear setting in Bechavod et al. (2022) no longer

---

[2]"Overqualified" means an applicant has skills significantly exceeding the requirement. He/she may fit more senior positions or desire a much higher salary than this position can offer.

hold, and simultaneously maximizing three types of welfare can be impossible. In the following sections, we further analyze the relationships between each pair of welfare in general non-linear settings.

4.2 Compatibility analysis of each welfare pair

Sec. 4.1 provides the necessary and sufficient condition for the alignment of all welfare. A natural question is whether there exist scenarios under which different pairs of welfare are compatible and can be simultaneously maximized. This section answers this question and explores the relationships between the welfare of every two parties in general settings.

Decision-maker welfare & agent welfare We first examine the relationships between *decision-maker welfare* $DW(f)$ and *agent welfare* $AW(f)$, the following result shows that the two can be maximized under the same f.

Proposition 4.4. *Under the realizability assumption that $h \in \mathcal{F}$, decision-maker welfare and agent welfare are compatible and can attain the optimum simultaneously.*

Note that *agent welfare* $AW(f) = 0$ as long as $f(x) \geq h(x)$ holds for all x. When $h \in \mathcal{F}$, the decision-maker can simply deploy $f = h$ to assign each agent the correct outcome; this ensures optimal welfare for both parties.

Decision-maker welfare & social welfare Unlike compatibility results in Sec 4.2, the *decision-maker welfare* and either element of *social welfare* (i.e., agent improvement $IMP(f)$ or agent safety $SF(f)$) in general are not compatible, and their compatibility needs all conditions in Thm. 4.1 to be satisfied.

Proposition 4.5. *The decision-maker welfare and social welfare are guaranteed to attain the optimum simultaneously regardless of the feature distribution P_X and the cost function $c(x, x')$ if and only if Thm. 4.1 holds.*

As illustrated in Example 4.2, each pair of welfare may be incompatible when $h \notin \mathcal{F}$. Intuitively, since the decision-maker cannot deploy $f = h$, it cannot make accurate decisions and provide a correct direction for agents to improve. Similarly, Example 4.3 shows the incompatibility when $\mathcal{F}$ is non-linear. Although the decision-maker can deploy $f = h$ to attain the optimal DW, the agents may fail to improve because their limited information level produces incorrect estimations of f. When $\mathcal{F}$ is non-linear, the gradient $\nabla f(x)$ is no longer a constant, and agents with $K = 1$ will respond in the wrong directions.

Agent improvement & agent safety While Prop. 4.5 holds for both measures (i.e., *agent improvement* $IMP(f)$ and *agent safety* $SF(f)$) of social welfare, the two in general are not compatible (as shown in Example 4.3). However, we can identify the necessary and sufficient conditions under which $IMP(f)$ and $SF(f)$ attain the maximum simultaneously in general non-linear settings.

Theorem 4.6 (Maximize agent improvement while ensuring safety). *Suppose agents have information level K and respond to f based on Def. 3.2. Let $f^* \in \arg\max_{f \in \mathcal{F}} IMP(f)$ be any policy that maximizes the agent improvement, and let*

$Q_x^{f^*}$ *be the estimation of f^* for agents with feature vector x. Then any f^* is guaranteed to be safe (i.e., $SF(f^*) = 0$) regardless of the feature distribution P_X and the cost function $c(x, x')$ if and only if **either** of the following scenarios holds: (i) Thm. 4.1 holds; (ii) For each agent with feature vector x, we have: $\frac{\partial h}{\partial \widetilde{x}_i} \cdot \frac{\partial Q_x^{f^*}}{\partial \widetilde{x}_i} \geq 0$ for any $\widetilde{x} = [\widetilde{x}_1, ..., \widetilde{x}_d] \in dom(X)$.*

It is trivial to see scenario *(i)* in Thm. 4.6 is a sufficient condition (it ensures the compatibility of three welfare). For scenario *(ii)*, it implies that for any arbitrary agent with feature x, the f^* will result in an estimation $Q_x^{f^*}$, under which each agent changes her features in directions to improve the qualification.

Social welfare & agent welfare Finally, we discuss the relationship between *social welfare* and *agent welfare*. Since *agent welfare* is optimal only when $f(x) \geq h(x)$ always holds, f satisfying such a constraint generally cannot maximize *social welfare*. In Example 4.2, a policy that maximizes agent improvement may underestimate certain agents' qualifications and lead to $AW(f) < 0$. Nonetheless, we can still find the necessary and sufficient conditions under which $IMP(f)$ and $SF(f)$ are compatible, as presented below.

Theorem 4.7 (Maximize social welfare and agent welfare). *Suppose agents have information level K and respond to f based on Def. 3.2. Denote $\mathcal{F}' = \{f : f \in \mathcal{F} \text{ and } f(x) \geq h(x), \forall x \in dom(X)\}$ as the set of policies maximizing the agent welfare. Then for any $f^* \in \arg\max_{f \in \mathcal{F}} IMP(f)$, f^* is guaranteed to also maximize $AW(f)$ regardless of the feature distribution P_X and the cost function $c(x, x')$ if and only if **either** of the following scenarios holds: (i) Thm. 4.1 holds; (ii) $\exists f' \in \mathcal{F}'$ and constant $C \in \mathbb{R}^+$ such that $Q_x^{f'} = Q_x^{f^*} + C$ for any $x \in dom(X)$.*

Again, scenario *(i)* in Thm. 4.7 is a sufficient condition, while scenario *(ii)* means that the decision-maker can find a classifier that induces perfect *agent welfare* while simultaneously giving the agents the same information as f^*. Note that $Q_x^{f'} = Q_x^{f^*} + C$ not necessarily implies $f' = f^* + C$ when $dom(X)$ is a zero-measure set. For example, when $K = 2$ and $dom(X) = \{0\}$, we can have $f^* = e^x - x - 1$ and $f' = \frac{x^2}{2} + 0.5$, where f^* belongs to the exponential function family but f' belongs to the quadratic function family.

To conclude, only *decision-maker welfare* and *agent welfare* can be maximized at the same time relatively easily, other pairs of welfare need strong conditions to align under the general settings. Even the two notions in *social welfare* (i.e., agent improvement and safety) can easily contradict each other, and the *decision-maker welfare* and *social welfare* need the most restrictive conditions (which is exactly Thm. 4.1) to be aligned. These results highlight the difficulties of balancing the welfare of different parties in general non-linear settings, which further motivates the "irreducible" optimization framework we will introduce in Sec. 5.

5 Welfare-Aware Optimization

In this section, we present our algorithm that balances the welfare of all parties in general settings. We formulate this welfare-aware learning problem under the strategic setting

Algorithm 1: Strategic Welfare-Aware Optimization (`STWF`)

Input: Training set $\mathcal{S}$, Agent response function Δ_ϕ, Agent information level K
Parameters: Total number of epochs n, batch size B, learning rate γ, hyperparameters λ_1, λ_2, initial parameters θ for f
Output: Classifier $f \in \mathcal{F}$

1: Train ground truth model $h \in \mathcal{H}$ with $\mathcal{S}$
2: **for** $t \in \{1, \dots n\}$ **do**
3:    **for** $k \in \{1, \dots \frac{n}{B}\}$ **do**
4:       Calculate $\widehat{Y}_B = f_\theta(X_B)$
5:       Calculate the post-response features of agents: $X_B^* = \Delta_\phi\big(X_B, I_k(f, x)\big)$
6:       Calculate the outcomes after responses: $Y_B^* = h(X_B^*)$
7:       Calculate $\ell_{\mathbf{DW}}, \ell_{\mathbf{SWF}}, \ell_{\mathbf{AW}}$
8:       Compute $L = \ell_{\mathbf{DW}} + \lambda_1 \cdot \ell_{\mathbf{SWF}} + \lambda_2 \cdot \ell_{\mathbf{AW}}$
9:       Update $\theta = \theta - \gamma \cdot \frac{\partial L}{\partial \theta}$
10:    **end for**
11: **end for**
12: Return θ

---

as a regularized optimization where each welfare violation is represented by a loss term. Formally, we write the optimization problem as follows.

$$\min_{f \in \mathcal{F}}\big\{\ell_{\mathbf{DW}} + \lambda_1 \cdot \ell_{\mathbf{SWF}} + \lambda_2 \cdot \ell_{\mathbf{AW}}\big\} \qquad (1)$$

where $\ell_{\mathbf{DW}}$, $\ell_{\mathbf{SWF}}$, $\ell_{\mathbf{AW}}$ are the losses corresponding to *decision-maker welfare*, *social welfare*, and *agent welfare*, respectively. The hyper-parameters $\lambda_1, \lambda_2 \geq 0$ balance the trade-off between the welfare of three parties. Given training dataset $\{(x_i, y_i)\}_{i=1}^N$, the decision-maker first learns h and then calculates each loss as follows:

- $\ell_{\mathbf{DW}}$: we may use common loss functions such as cross-entropy loss, 0-1 loss, etc., and quantify $\ell_{\mathbf{DW}}$ as the average loss over the dataset. In experiments (Sec. 6), we adopt $\ell_{\mathbf{DW}} = \frac{1}{N} \sum_{i=1}^N -\big(y_i \log\big(f(x_i)\big) + (1 - y_i) \log\big(1 - f(x_i)\big)\big)$.

- $\ell_{\mathbf{SWF}}$: since both *agent improvement* and *agent safety* indicate the *social welfare*. We define $\ell_{\mathbf{SWF}} = \ell_{\mathbf{IMP}} + \ell_{\mathbf{SF}}$, where $\ell_{\mathbf{IMP}}$ is the loss associated with agent improvement and $\ell_{\mathbf{SF}}$ measures the violation of agent safety. In experiments (Sec. 6), we also present empirical results when $\ell_{\mathbf{SWF}}$ only includes one of the two losses. Specifically, the two losses are defined as follows.

  - $\ell_{\mathbf{IMP}}$: it aims to penalize f when it induces less agent improvement. We can write it as the cross entropy loss over $h(x_i^*)$ (i.e., the qualification after agent response) and 1 (i.e., perfectly qualified). Formally, $\ell_{\mathbf{IMP}} = \frac{1}{N} \sum_{i=1}^N -\log(h(x_i^*))$.
  - $\ell_{\mathbf{SF}}$: it penalizes the model whenever the agent qualification deteriorates after the response. Formally, $\ell_{\mathbf{SF}} = \frac{1}{N} \sum_{x \in \mathcal{A}_{SF}} -\log(h(x^*))$ where the set $\mathcal{A}_{SF} = \{x_i | h(x_i^*) < h(x_i)\}$ includes all agents with deteriorated qualifications.

- $\ell_{\mathbf{AW}}$: it focuses on *agent welfare* and penalizes f whenever the model underestimates the actual qualification of the agent. Formally, $\ell_{\mathbf{AW}} = \frac{1}{N} \sum_{x \in \mathcal{A}_{AW}} -\log(f(x))$

where the set $\mathcal{A}_{AW} = \{x_i | f(x_i) < h(x_i)\}$ includes all agents that are underestimated by the model.

The optimizer of (1) is Pareto optimal and it can be solved using gradient-based algorithms, as long as the gradient of losses for the model parameter (denoted as θ) exists. Since f, h are continuous, gradients already exist for $\ell_{\mathbf{DW}}$ and $\ell_{\mathbf{AW}}$. For $\ell_{\mathbf{SWF}}$, because it is a function of post-response feature X^*, we need gradients to exist for X^* as well. If we already know how agents respond to the model (i.e., X^* as a function of f illustrated in Def. 3.2), then we only need the $K + 1^{th}$ gradients to exist for $f(x)$ with respect to θ. When $K = 1$, the objective (1) can be easily optimized if $\nabla f(x)$ is continuous in θ, which is a relatively mild requirement and has also been used in previous literature (Rosenfeld et al. 2020). The complete algorithm for <u>S</u>tra<u>T</u>egic <u>W</u>el<u>F</u>are-Aware optimization (`STWF`) is shown in Alg. 1.

Learning agent response function Δ_ϕ. Note that while Sec. 3 introduces agent response following Def. 3.2, our algorithm does not have this requirement; In practice, agent behavior can be more complicated, and we can consider a general learnable response function $x^* = \Delta_\phi(x, I_k(f, x))$. Note that Levanon and Rosenfeld (2021) proposed a concave optimization protocol to learn a response function linear in any representation space of agents' features. However, their design implicitly assumes the agents have the perfect knowledge of the representation mapping function used by the decision-maker so that they respond based on their representations, which is unrealistic. Here, we avoid this by only assuming the response is a parameterized function of agents' features x and the information up to level K, i.e., $x^* = \Delta_\phi(x, I_k(f, x))$ where ϕ are parameters for some models such as neural networks and $I_k(f, x)$ is all available information for agents with information level K and feature x. With this assumption, we provide Alg. 2 as a general protocol to learn ϕ. Then we can plug the learned Δ_ϕ into `STWF` (Alg. 1) to optimize the welfare.

Algorithm 2: AgentResponse

Input: Experimental training set $\mathcal{S}$ containing n samples $\{x_j, y_j\}_{j=1}^n$, N arbitrary classifiers $\{f_i\}_{i=1}^N$, ground-truth model h, agent information level K
Parameters: initial parameters ϕ for the agent response function $\Delta_\phi(x, I_k(f, x))$
Output: Parameters ϕ

1: $\mathcal{S}_0 \leftarrow \{\}$
2: **for** $i \in \{1, \dots N\}$ **do**
3:    Impose f_i on $\mathcal{S}$ and obtain x^* where x^* is the post-response feature vector
4:    $\mathcal{S}_i \leftarrow \{\{x_j, I_k(f_i, x_j), x^*\}\}$ {build up the new training set for agent response}
5:    $\mathcal{S} = \mathcal{S}_{i-1} \cup \mathcal{S}_i$
6: **end for**
7: Train Δ_ϕ with $\mathcal{S}$
8: Return ϕ

Alg. 2 requires a population for controlled experiments. As pointed out by Miller, Milli, and Hardt (2020), we can never learn the agent response without the controlled trials. Though the availability of controlled experiments has

been assumed in most literature on *performative prediction* (Perdomo et al. 2020; Izzo, Ying, and Zou 2021), it may be costly to conduct the experiments in practice.

6 Experiments

We conduct comprehensive experiments on one synthetic and two real-world datasets to validate STWF[3]. We first illustrate the trade-offs for each pair of welfare and examine how they can be balanced by adjusting λ_1, λ_2 in Alg. 1. Then, we compare STWF with previous algorithms that only consider the welfare of a subset of parties. Since many of the existing algorithms also consider fairness across agents from different groups, we also evaluate the unfairness induced by our algorithm in addition to decision-maker, agent, and social welfare. We use a three-layer ReLU neural network to learn labeling function h and logistic regression to train the policy f, and $D(x) = \mathbf{1}(f(x) \geq 0.5)$. For each parameter setting, we perform 20 independent runs of the experiments with different random seeds and report the average.

6.1 Datasets

Next, we introduce the data used in the experiments. For all datasets, 80% of the data is used for training.

- **Synthetic data.** It has two features $X = (X_1, X_2)$ and a sensitive attribute $Z \in \{0, 1\}$ for demographic groups. Let $(X|Z = z)$ be multivariate Gaussian random variables with different means and variances conditioned on Z. Labeling function $h(x)$ is a quadratic function in both x_1, x_2, and the label $Y = \mathbf{1}(h(X) \geq 0.5)$. We generate 10000 independent samples with 20% of them having $Z = 1$ and the rest $Z = 0$. We assume all agents have the same cost function $c(x, x') = 5 \cdot \|x' - x\|^2$.

- **ACSIncome-CA data** (Ding et al. 2021). It contains income-related data for California residents. Let label Y be whether an agent's annual income is higher than 50K USD and let the sensitive attribute Z be sex. We randomly select 20000 samples from the data of year 2018. We assume all agents have the same effort budget $c(x, x') = \|x' - x\|^2$.

- **German Credit data** (Dua and Graff 2017). It has 1000 samples and the label Y is whether the agent's credit risk is high or low. We adopt the data preprocessed by (Jiang and Nachum 2020) where age serves as the sensitive attribute to split agents into two groups. We assume the agents are able to improve the following features: existing account status, credit history, credit amount, saving account, present employment, installment rate, guarantors, residence. We assume all agents have the same effort budget $c(x, x') = \|x' - x\|^2$.

6.2 Balance the Welfare with Different λ_1, λ_2

We first validate STWF and show that adjusting the hyperparameters λ_1, λ_2 effectively balances the welfare of different parties. The results in Fig. 3 are as expected, where *social welfare* SWF(f) increases in λ_1 and *agent welfare* AW(f)

---

[3]https://github.com/osu-srml/StrategicWelfare

increases in λ_2. This shows the effectiveness of using our algorithm to adjust the allocation of welfare under different situations. As we analyzed in Sec. 4, under the general non-linear setting, each pair of welfare possibly conflicts. Fig. 3 validates this by demonstrating the trade-offs between different welfare pairs on two datasets, while it also reveals that the conflicts are not always serious between each pair. For the synthetic data, increasing *social welfare* slightly conflicts with *decision-maker welfare*, while increasing *agent welfare* drastically conflicts with the *decision-maker welfare*. But the trade-off between *social welfare* and *agent welfare* is not obvious. For the German Credit dataset, increasing *social welfare* seems to slightly conflict with both *agent welfare* and *decision-maker welfare*, while increasing *agent welfare* has an obvious trade-off with *decision-maker* welfare. Conversely, for the ACSIncome-CA dataset, increasing *social welfare* only sacrifices *agent welfare*, but increasing *agent welfare* incurs conflicts with both other welfare.

6.3 Comparison with Previous Algorithms

Since STWF is the first optimization protocol that simultaneously considers the welfare of the decision-maker, agent, and society, there exists no algorithm that we can directly compare with. Nonetheless, we can adapt existing algorithms that only consider the welfare of a subset of parties to our setting, and compare them with STWF.

Specifically, we compare STWF with four algorithms:

- **Empirical Risk Minimization (ERM):** It only considers *decision-maker welfare* by minimizing the predictive loss on agent current data.

- **Safety (SAFE)** (Rosenfeld et al. 2020): It considers both *decision-maker welfare* and *agent safety*. The goal is to train an accurate model without hurting agent qualification.

- **Equal Improvement (EI)** (Guldogan et al. 2022): It considers *agent improvement* by equalizing the probability of unqualified agents becoming qualified from different groups, as opposed to maximizing agent improvement considered in our work.

- **Bounded Effort (BE)** (Heidari, Nanda, and Gummadi 2019): It considers *agent improvement* by equalizing the proportion of new agents becoming qualified after the best response. The difference between BE and EI is a bit subtle, where we show details in Tab. 2.

Note that both EI and BE focus on fairness across different groups. To comprehensively compare different algorithms, we also evaluate the unfairness induced by each algorithm. Specifically, we consider both fairness notions with and without considering agent behavior, including Equal Improvability (EI), Bounded Effort (BE), Demographic Parity (DP), and Equal Opportunity (EO), which are defined in Tab. 2. We run all experiments on a MacBook Pro with Apple M1 Pro chips, memory of 16GB, and Python 3.9.13. In all experiments of Sec. 6.3, we use the ADAM optimizer. Except ERM, we first perform hyper-parameter selections for all algorithms, e.g., we select λ_1, λ_2 of STWF by grid searching with cross-validation. For STWF, we perform cross-validation on 7 different seeds to select the learning rate

Table 1: Comparisons of welfare between STWF and other benchmark algorithms where the largest values are boldfaced.

DATASET	METRIC	ALGORITHMS				
		ERM	STWF	SAFE	EI	BE
SYNTHETIC	TOTAL	0.76 ± 0.08	$\mathbf{1.08} \pm .16$	0.74 ± 0.09	0.76 ± 0.08	0.75 ± 0.10
	DW	0.71 ± 0.05	$\mathbf{0.75} \pm 0.08$	0.71 ± 0.05	0.71 ± 0.06	0.71 ± 0.05
	SWF	0.09 ± 0.05	$\mathbf{0.3} \pm 0.08$	0.07 ± 0.06	0.09 ± 0.05	0.08 ± 0.07
	AW	-0.22 ± 0.001	$\mathbf{-0.14} \pm 0.001$	-0.22 ± 0.001	-0.23 ± 0.001	-0.23 ± 0.001
GERMAN CREDIT	TOTAL	$\mathbf{0.710} \pm 0.10$	$0.704 \pm .14$	0.703 ± 0.11	0.707 ± 0.11	0.706 ± 0.11
	DW	$\mathbf{0.781} \pm 0.10$	0.736 ± 0.13	0.779 ± 0.10	0.780 ± 0.10	0.777 ± 0.10
	SWF	0.030 ± 0.03	0.022 ± 0.04	0.030 ± 0.04	$\mathbf{0.031} \pm 0.04$	0.030 ± 0.04
	AW	-0.100 ± 0.00	$\mathbf{-0.052} \pm 0.00$	-0.099 ± 0.00	-0.100 ± 0.00	0.100 ± 0.00
ACSIncome-CA	TOTAL	0.74 ± 0.06	$\mathbf{0.95} \pm .15$	0.74 ± 0.04	0.74 ± 0.06	0.73 ± 0.08
	DW	0.76 ± 0.03	$\mathbf{0.77} \pm 0.05$	0.76 ± 0.03	0.76 ± 0.13	0.76 ± 0.05
	SWF	0.12 ± 0.05	$\mathbf{0.40} \pm 0.10$	0.12 ± 0.05	0.12 ± 0.04	0.12 ± 0.05
	AW	-0.16 ± 0.00	-0.21 ± 0.00	$\mathbf{-0.15} \pm 0.00$	-0.16 ± 0.00	-0.16 ± 0.00

(a) Welfare while increasing λ_1 and keeping $\lambda_2 = 0$.

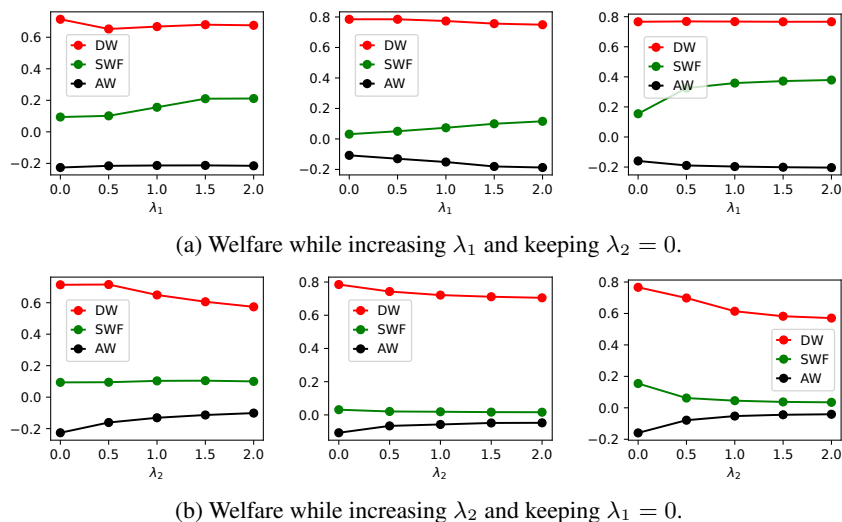(b) Welfare while increasing λ_2 and keeping $\lambda_1 = 0$.

Figure 3: Effects of adjusting λ_1, λ_2 on the welfare of each party: synthetic data (left plot), German Credit data (middle plot), ACSIncome-CA data (right plot). Increasing λ_1 (λ_2) results in improved SWF (AW) and trade-offs exist.

from $\{0.001, 0.01, 0.1\}$ and $\lambda_1, \lambda_2 \in \{0, 0.5, 1, 1.5, 2\}$. For EI, BE and SAFE, we choose the hyperparameters in the same way as the previous works (Guldogan et al. 2022; Rosenfeld et al. 2020; Heidari, Nanda, and Gummadi 2019). Specifically, λ (the strength of the regularization) is 0.1 for all these algorithms.

Comparison of welfare. Tab. 1 summarizes the performances of all algorithms, where STWF produces the largest total welfare on the synthetic dataset and the ACSIncome-CA dataset. For the German Credit dataset, all algorithms produce similar welfare, while the ERM has a slight advantage. More clearly, as shown in the left plot from Fig. 4a to Fig. 4c, STWF gains a large advantage, demonstrating the potential benefits of training ML model in a welfare-aware manner. Meanwhile, although STWF does not have a similar advantage on the German Credit dataset, it produces a much more balanced welfare allocation, suggesting the ability of STWF to take care of each party under the strategic learning setting. However, since we select λ_1, λ_2 only based on the sum of the welfare, it is possible to sacrifice the welfare of some specific parties (e.g., *agent welfare* in the ACSincome-

CA dataset). This can be solved by further adjusting λ_1, λ_2 based on other selection criteria.

Besides, we also illustrate that safety is not automatically guaranteed if we only aim to maximize agent improvement in the synthetic dataset. Specifically, if we let $L_{SWF} = L_{IMP}$, we can visualize the safety in Fig. 5. It is obvious that only when L_{SF} is added into L_{SWF} will the decision rule become perfectly safe as λ_1 increases.

Improve welfare does not ensure fairness. The middle and the right plots of Fig. 4 compare the unfairness of different algorithms. Although STWF has the highest welfare, it does not ensure fairness. These results further shed light on a thought-provoking question: how likely is it for a fairness-aware algorithm to "harm" each party's welfare under the strategic setting? Recent literature on fairness (Guldogan et al. 2022; Heidari, Nanda, and Gummadi 2019) already began to consider "agent improvement" when strategic behaviors are present. However, it remains an intriguing direction to consider fairness and welfare together.
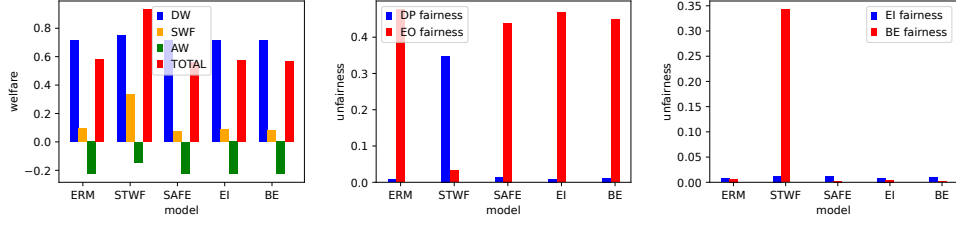
(a) Comparisons under the synthetic dataset.
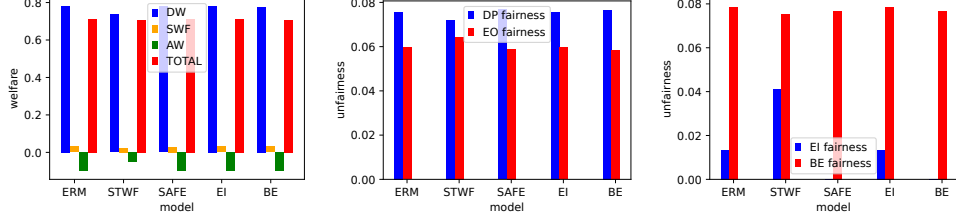
(b) Comparisons under the German Credit dataset.
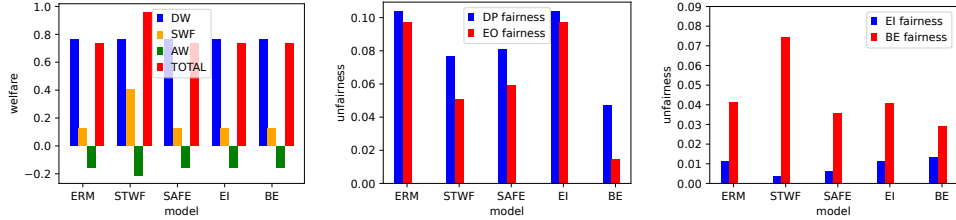
(c) Comparisons under the ACSincome-CA dataset.

Figure 4: Comparisons of welfare (left plot) and unfairness (middle & right plot) for different algorithms.

Table 2: Fairness notions used in our experiments.

Notion	Definition
Equal Improvability (EI)	$P\left(f(x^*) \geq 0.5 \mid f(x) < 0.5, Z = z\right) = P\left(f(x^*) \geq 0.5 \mid f(x) < 0.5\right)$
Bounded Effort (BE)	$P\left(f(x^*) \geq 0.5, f(x) < 0.5 \mid Z = z\right) = P\left(f(x^*) \geq 0.5, f(x) < 0.5\right)$
Demographic Parity (DP)	$P\left(f(x) \geq 0.5 \mid Z = z\right) = P\left(f(x) \geq 0.5\right)$
Equal Opportunity (EO)	$P\left(f(x) \geq 0.5 \mid Y = 1, Z = z\right) = P\left(f(x) \geq 0.5 \mid Y = 1\right)$
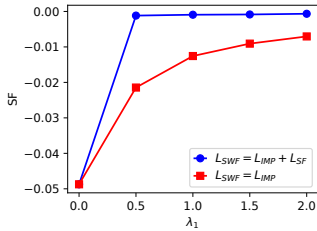
Figure 5: Comparisons of safety under the synthetic dataset.

highly complex. With incorrect estimations of agents' behavior, our algorithm may fail and lead to unexpected outcomes. Thus, it is worthwhile for future work to collect real data from human feedback, and learn agent response function. Additionally, our experiments demonstrate that welfare and fairness may not imply each other, making it necessary to consider both notions theoretically to further promote trustworthy machine learning.

7 Conclusion & Societal Impacts

To facilitate socially responsible decision-making on strategic human agents in practice, our work is the first to consider the welfare of all parties in strategic learning under non-linear settings when the agents have specific information levels. Our algorithm relies on the accurate estimation of agent response. In practice, the agent's behavior can be

References

Ahmadi, S.; Beyhaghi, H.; Blum, A.; and Naggita, K. 2021. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, 6–25.

Ahmadi, S.; Beyhaghi, H.; Blum, A.; and Naggita, K. 2022a. On classification of strategic agents who can both game and improve. *arXiv preprint arXiv:2203.00124*.

Ahmadi, S.; Beyhaghi, H.; Blum, A.; and Naggita, K. 2022b. Setting Fair Incentives to Maximize Improvement. *arXiv preprint arXiv:2203.00134*.

Alon, T.; Dobson, M.; Procaccia, A.; Talgam-Cohen, I.; and Tucker-Foltz, J. 2020. Multiagent Evaluation Mechanisms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 1774–1781.

Barsotti, F.; Kocer, R. G.; and Santos, F. P. 2022. Transparency, Detection and Imitation in Strategic Classification. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 67–73.

Bechavod, Y.; Ligett, K.; Wu, S.; and Ziani, J. 2021. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*, 1234–1242.

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

Ben-Porat, O.; and Tennenholtz, M. 2017. Best Response Regression. In *Advances in Neural Information Processing Systems*.

Braverman, M.; and Garg, S. 2020. The Role of Randomness and Noise in Strategic Classification. *CoRR*, abs/2005.08377.

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. 2020. Strategic Recourse in Linear Classification. *CoRR*, abs/2011.00355.

Dean, S.; Curmei, M.; Ratliff, L. J.; Morgenstern, J.; and Fazel, M. 2023. Emergent segmentation from participation dynamics and multi-learner retraining. arXiv:2206.02667.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34: 6478–6490.

Dong, J.; Roth, A.; Schutzman, Z.; Waggoner, B.; and Wu, Z. S. 2018. Strategic Classification from Revealed Preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 55–70.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository.

Eilat, I.; Finkelshtein, B.; Baskin, C.; and Rosenfeld, N. 2022. Strategic Classification with Graph Neural Networks.

Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway feedback loops in predictive policing. In *Conference on fairness, accountability and transparency*, 160–171. PMLR.

Estornell, A.; Das, S.; and Vorobeychik, Y. 2021. Incentivizing Truthfulness Through Audits in Strategic Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6, 5347–5354.

Florio, M. 2014. *Applied welfare economics: Cost-benefit analysis of projects and policies*. Routledge.

Guldogan, O.; Zeng, Y.; Sohn, J.-y.; Pedarsani, R.; and Lee, K. 2022. Equal Improvability: A New Fairness Notion Considering the Long-term Impact. In *The Eleventh International Conference on Learning Representations*.

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. *Advances in Neural Information Processing Systems*, 28728–28741.

Heidari, H.; Nanda, V.; and Gummadi, K. 2019. On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning. In *36th International Conference on Machine Learning*, 2692–2701.

Horowitz, G.; and Rosenfeld, N. 2023. Causal Strategic Classification: A Tale of Two Shifts. arXiv:2302.06280.

Izzo, Z.; Ying, L.; and Zou, J. 2021. How to Learn when Data Reacts to Your Model: Performative Gradient Descent. In *Proceedings of the 38th International Conference on Machine Learning*, 4641–4650.

Jagadeesan, M.; Mendler-Dünner, C.; and Hardt, M. 2021. Alternative Microfoundations for Strategic Classification. In *Proceedings of the 38th International Conference on Machine Learning*, 4687–4697.

Jiang, H.; and Nachum, O. 2020. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, 702–712. PMLR.

Jin, K.; Xie, T.; Liu, Y.; and Zhang, X. 2024. Addressing Polarization and Unfairness in Performative Prediction. *arXiv preprint arXiv:2406.16756*.

Jin, K.; Zhang, X.; Khalili, M. M.; Naghizadeh, P.; and Liu, M. 2022. Incentive Mechanisms for Strategic Classification and Regression Problems. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, 760–790.

Kleinberg, J.; and Raghavan, M. 2020. How Do Classifiers Induce Agents to Invest Effort Strategically? 1–23.

Lechner, T.; Urner, R.; and Ben-David, S. 2023. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*, 18714–18732. PMLR.

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

Levanon, S.; and Rosenfeld, N. 2022. Generalized Strategic Classification and the Case of Aligned Incentives. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, 12593–12618.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2019. Delayed Impact of Fair Machine Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 6196–6200.

Liu, L. T.; Wilson, A.; Haghtalab, N.; Kalai, A. T.; Borgs, C.; and Chayes, J. 2020. The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 381–391.

Miller, J.; Milli, S.; and Hardt, M. 2020. Strategic Classification is Causal Modeling in Disguise. In *Proceedings of the 37th International Conference on Machine Learning*.

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.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, 7599–7609.

Raab, R.; and Liu, Y. 2021. Unintended selection: Persistent qualification rate disparities and interventions. *Advances in Neural Information Processing Systems*, 26053–26065.

Rolf, E.; Simchowitz, M.; Dean, S.; Liu, L. T.; Bjorkegren, D.; Hardt, M.; and Blumenstock, J. 2020. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. In *International Conference on Machine Learning*, 8158–8168. PMLR.

Rosenfeld, N.; Hilgard, A.; Ravindranath, S. S.; and Parkes, D. C. 2020. From Predictions to Decisions: Using Lookahead Regularization. In *Advances in Neural Information Processing Systems*, 4115–4126.

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. L.; and Axelrod, B. 2020. Causal Strategic Linear Regression. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20.

Sundaram, R.; Vullikanti, A.; Xu, H.; and Yao, F. 2021. PAC-Learning for Strategic Classification. In *Proceedings of the 38th International Conference on Machine Learning*, 9978–9988.

Tang, W.; Ho, C.-J.; and Liu, Y. 2021. Linear Models are Robust Optimal Under Strategic Behavior. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 2584–2592.

Xie, T.; and Zhang, X. 2024. Automating Data Annotation under Strategic Human Agents: Risks and Potential Solutions. *arXiv preprint arXiv:2405.08027*.

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.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? In *Advances in Neural Information Processing Systems*, 18457–18469.

A Proofs

A.1 Proof of Prop. 3.3

Proof. (i) Proof of the necessity: The fact that agents are guaranteed to best respond means they can estimate f perfectly. Since the agents have information level K, they will estimate $f(x)$ by the Taylor expansion of $f(x)$ at x using gradients up to the Kth order and this results in a polynomial function with order at most K as specified in Eqn. (2). Note that if the gradient of any order does not exist, the Taylor expansion will definitely not yield a correct estimation of f.

$$f(x') = f(x) + \sum_{i=1}^{d} \frac{\partial f}{\partial x_i}(x)(x'_i - x_i) + \frac{1}{2!}\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2 f}{\partial x_i \partial x_j}(x)(x'_i - x_i)(x'_j - x_j) + \cdots \tag{2}$$

$$+ \frac{1}{K!}\sum_{i_1=1}^{d}\cdots\sum_{i_K=1}^{d} \frac{\partial^K f}{\partial x_{i_1}\cdots\partial x_{i_K}}(x)(x'_{i_1} - x_{i_1})\cdots(x'_{i_K} - x_{i_K})$$

(ii) Proof of sufficiency: When f is a polynomial function with order $k \geq K$, the agents with information level K are guaranteed to know all levels of gradients of f which are non-zero. Then the agents can construct f perfectly according to Eqn. (2), then the agent response according to Def. 3.2 is indeed the best response. $\square$

A.2 Proof of Thm. 4.1

Proof. 1. Proof of the necessity: Since the maximization of the decision-maker welfare regardless of P_X needs $f = h$ (otherwise, one can always take the point mass on the point that f misidentifies as the new feature distribution and f is no longer the one maximizing the *decision-maker welfare*), then $h \in \mathcal{F}$ is needed. Then the agents need to best respond to f to ensure the maximization of social welfare . If there is an arbitrary agent who does not best respond to $f = h$, this means she misses the largest possible improvement and the overall maximization of social welfare is failed. Thus, according to Prop. 3.3, h itself must be a polynomial of X with at most order K to guarantee agents to best respond.

2. Proof of sufficiency: When h is a polynomial with order at most K and $h \in \mathcal{F}$, the decision-maker can directly deploy $h = f$ to ensure the maximization of all welfare. $\square$

A.3 Proof of Prop. 4.5

Proof. 1. Proof of the necessity: Since the maximization of the decision-maker welfare needs $f = h$, then $h \in \mathcal{F}$ is needed. Then the agents need to best respond to f to ensure the maximization of social welfare. If there is an arbitrary agent who does not best respond to $f = h$, this means she misses the largest possible improvement and the overall maximization of social welfare is failed. Thus, according to Prop. 3.3, h itself must be a polynomial of X with at most order K to guarantee agents to best respond.

2. Proof of sufficiency: Thm. 4.1 is already a sufficient condition to guarantee the compatibility of *decision-maker welfare* and *social welfare*. $\square$

A.4 Proof of Thm. 4.6

Proof. 1. Proof of the necessity: When (i) does not hold, f^* is not equal to h and the agents cannot perfectly estimate f^* with $Q_x^{f^*}$. Therefore, if (ii) does not hold, this means there must exist a cost function $c(x, x')$, under which we can find a region $A \subset \text{dom}(X)$ where $Q_x^{f^*}$ increases while h decreases inside A. Thus, if we specify a cost function and an agent to satisfy the requirement that the agent can only change her features within A and when P_X is any continuous distribution, then f^* is unsafe, and we prove that either (i) or (ii) should at least hold.

2. Proof of the sufficiency: (i) is already a sufficient condition. Then for (ii), we know for each dimension i, $Q_x^{f^*}$ improves at the same direction as h and only the magnitude differs. This means agents will never be misguided by $Q_x^{f^*}$, thereby ensuring the safety. $\square$

A.5 Proof of Thm. 4.7

Proof. 1. Proof of the necessity: When (i) does not hold, f^* is not equal to h and the agents cannot perfectly estimate f^* with $Q_x^{f^*}$. Therefore, if (ii) also does not hold, this means any classifier f' maximizing the *agent welfare* does not let the agents get an estimation $Q_x^{f'}$ only differing in constant to $Q_{imp,x}$ which is the estimation for f^*. Thus, we can find a region $A \subset \text{dom}(X)$ where $Q_x^{f^*}$ where the $\max_A Q_x^{f^*} - \min_A Q_x^{f^*}$ does not equal to $\max_A Q_x^{f'} - \min_A Q_x^{f'}$. Then if we find an agent and a cost function to let the agent can only modify her features within A, her improvement under f' will be different from her improvement under f^*, which means f' cannot maximize *social welfare*.
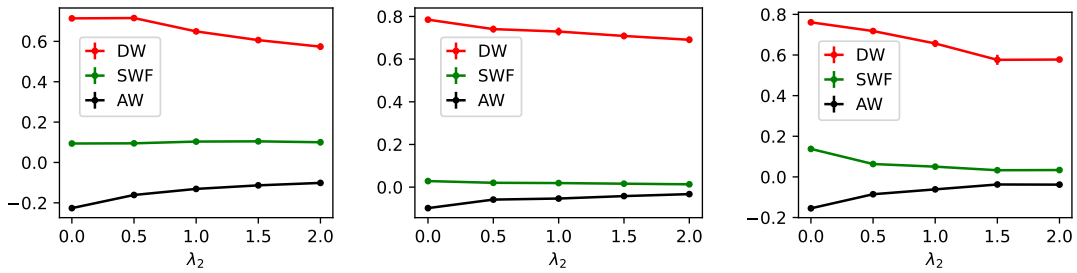
2. Proof of the sufficiency: (i) is already a sufficient condition. Then for (ii), we know $Q_x^{f'}$ induces the same amount of improvement as $Q_x^{f^*}$, which completes the proof. $\square$

B   Plots with error bars

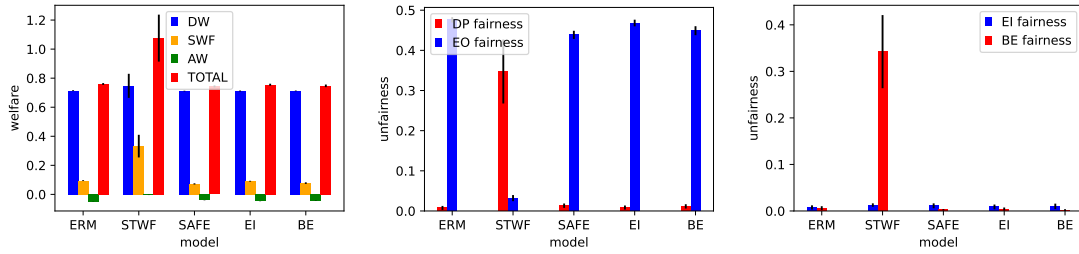We provide the error bar version of the previous figures as follows (Fig. 6, Fig. 7).
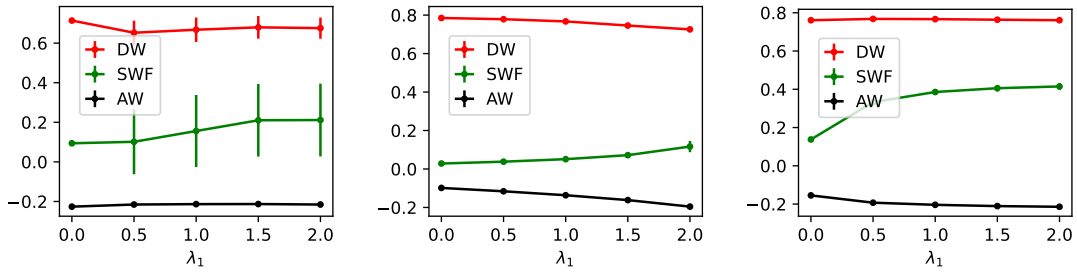
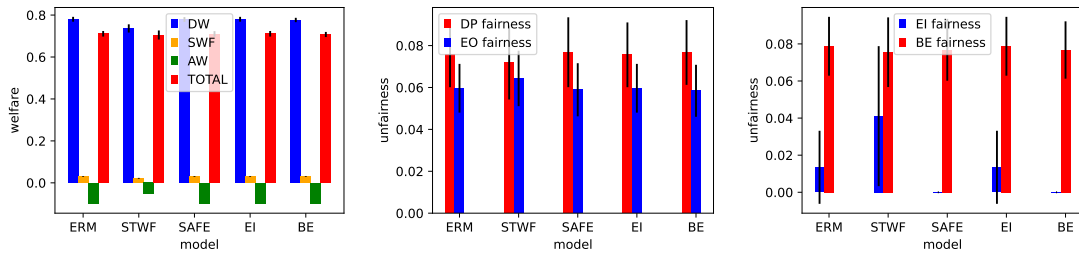(a) Dynamics of welfare while increasing λ_1 and keeping $\lambda_2 = 0$ in Eqn.(1).

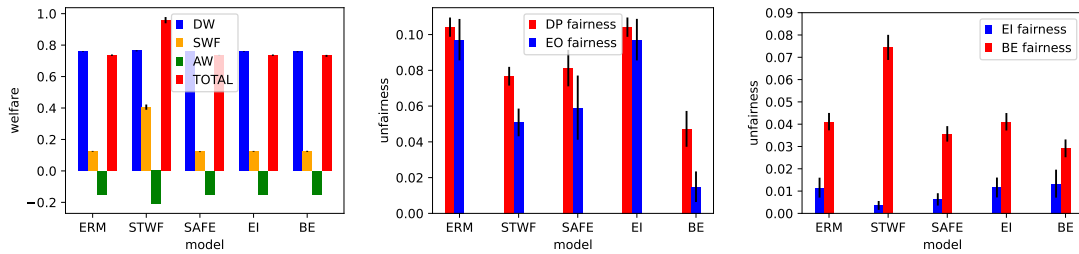(b) Dynamics of welfare while increasing λ_2 and keeping $\lambda_1 = 0$ in Eqn.(1).

Figure 6: Error bar version of Fig. 3

(a) Comparisons under the synthetic dataset.

(b) Comparisons under the German Credit dataset.

(c) Comparisons under the ACSincome-CA dataset.

Figure 7: Error bar version of Fig. 4.