To Give or Not to Give? The Impacts of Strategically Withheld Recourse

Yatong Chen
MPI for Intelligent Systems,
Tübingen AI Center, Tübingen, Germany

Andrew Estornell
Bytedance Research

Yevgeniy Vorobeychik
Washington University in Saint Louis

Yang Liu
University of California, Santa Cruz

Abstract

Individuals often aim to reverse undesired outcomes in interactions with automated systems, like loan denials, by either implementing system-recommended actions (recourse), or manipulating their features. While providing recourse benefits users and enhances system utility, it also provides information about the decision process that can be used for more effective strategic manipulation, especially when the individuals collectively share such information with each other. We show that this tension leads rational utility-maximizing systems to frequently withhold recourse, resulting in decreased population utility, particularly impacting sensitive groups. To mitigate these effects, we explore the role of recourse subsidies, finding them effective in increasing the provision of recourse actions by rational systems, as well as lowering the potential social cost and mitigating unfairness caused by recourse withholding.

1 INTRODUCTION

When individuals interacting with automated systems are denied a desired outcome (e.g., loan approval), they may seek a means of reversing this decision to obtain the desired outcome. This procedure is commonly referred to as *recourse* (Ustun et al., 2019). In cases where the system's decision rule is opaque (e.g.,

lending), the system itself is responsible for supplying individuals with recourse, i.e., a recommended feature modification that is feasible and will result in that individual being approved.

When the feature modification changes an agent's true qualification rate (e.g., paying off debt increases one's creditworthiness), providing recourse can benefit the system. However, offering recourse actions also exposes information about the system's decision rule, as each action leads to a positively classified feature vector close to the decision boundary. This added transparency creates opportunities for strategic individuals to exploit the system's decision rule by manipulating their features, especially when they share their knowledge about the decision rule with one another. For example, platforms like *GradCafe* for graduate school admissions and *LendingClub* for loan applications allow agents to see other applicants' features. This enables them to potentially *misreport* their features to mimic those of others, thereby leveraging publicly available information to their advantage (Bechavod et al., 2022; Chen et al., 2020; Estornell et al., 2023b; Hardt et al., 2016; Vorobeychik, 2023). Such feature manipulation can often reduce both system and social utility since it will increase the false positive rate. This creates a tension in providing recourse, where the utility gained from increased qualifications must be balanced against the utility lost due to manipulation that exploits the counterfactual information in recourse recommendations The consequence of this tension is that in many settings, providing recourse to all, or even most, of the agents may be suboptimal from a system's perspective. This sharply contrasts with the common assumption in the algorithmic recourse literature, which typically considers agents taking recourse actions without the possibility of manipulation.

On the other hand, we can consider subsidies as a

means to incentivize systems to offer recourse actions. Subsidy (Hu et al., 2019), or government incentive, is a type of government expenditure to financially help individuals, households, and businesses in various settings. Consider the *small business administration (SBA) microloan program* [1] in the United States as a motivating example. This program provides small loans to startups and small businesses and offers technical assistance and financial training to help borrowers succeed. In this work, we model subsidies that lower each individual's recourse costs, requiring the agent to pay only a fraction of the original amount. Compared with using penalty to disincentivize manipulation (Blocki et al., 2013) or using auditing to incentivize recourse taking (Estornell et al., 2023a), the main benefit of subsidies is that it requires no verification power from the system, reducing the potential harm caused by unintentionally impose large fines on truthful agents. We add a more detailed discussion in Section 2.

High-Level Overview of Our Model There are two parties in our setting: a utility-maximizing recourse system and a set of agents. Each agent is represented by a feature vector $\mathbf{x} \in \mathcal{X}$. The system trained a *fixed*, potentially opaque function $f : \mathcal{X} \rightarrow [0, 1]$ to decide who to provide a resource (e.g., loan) based on $\mathbf{x}$. For negatively classified agents, the system decides whether to provide a recourse action or not. The central tension comes from the fact that agents can both (1) lie about their features and manipulate them to some publicly known positively classified features and (2) take the recommended recourse actions that change their true features. Only the latter leads to an increase in the system's utility. The publicly known features come from either agents who are already classified positively, or agents who successfully obtain a recourse action from the system. The latter is more within the system's control and could be an easier target for manipulation, as they are more likely to be closer to the decision boundary. Thus, the system's main tool is to strategically withhold recourse actions from some agents to maximize their utility. Based on the relative cost of recourse and manipulation, agents choose to take the recommended action or manipulate known positively classified features. See Figure 1 for a demonstration of our modeling framework.

Main Results We show that in many cases, the system is incentivized to strategically withhold recourse from most if not all, agents to prevent manipulation. To our knowledge, this is the first work to challenge the assumption that a utility-maximizing recourse system

will naturally provide recourse without third-party intervention (e.g., government regulation). As fewer agents receive recourse, the *social cost*—the average cost to achieve positive classification—rises. Withholding recourse also limits legitimate paths to positive classification, pushing more individuals toward manipulation. This burden often falls disproportionately on disadvantaged groups, worsening existing inequalities. To address this, we explore recourse subsidies, a third-party payment that reduces recourse costs, and find them effective in increasing recourse providing, reducing social costs, and mitigating unfairness.

The details for reproducing our experimental results can be found at https://github.com/UCSC-REAL/Strategic-withheld-recourse.

2 RELATED WORKS

Our work is closely related to the literature on algorithmic recourse, strategic classification, and fairness in general. Due to the page limit, additional related work on fairness and social cost in strategic classification and recourse (Gupta et al., 2019; von Kügelgen et al., 2022; Ehyaei et al., 2023; Estornell et al., 2023b), transparency (Barsotti et al., 2022; Akyol et al., 2016) and others can be found in Appendix B.

Recourse Much of the line of algorithmic recourse (Ustun et al., 2019; Venkatasubramanian & Alfano, 2020; Karimi et al., 2020a; Gupta et al., 2019; Karimi et al., 2020b; von Kügelgen et al., 2020; Chen et al., 2020; Harris et al., 2022) focuses on the setting where the requested recourse is guaranteed to be provided out of ethical consideration (Venkatasubramanian & Alfano, 2020). Our work is the first to challenge this fundamental assumption and argue that without a third-party's intervention, a utility-maximizing algorithmic recourse system may be incentivized to withhold recourse from some agents to prevent manipulations strategically. We point the reader to Karimi et al. (2020a) for a more detailed discussion of the concepts and recent development of algorithmic recourse.

Strategic Classification Strategic classification focuses on the problem of how to effectively make predictions in the presence of agents who behave strategically to obtain desirable outcomes (Hardt et al., 2016; Chen et al., 2018; Tsirtsis et al., 2019; Levanon & Rosenfeld, 2021; Dong et al., 2018; Chen et al., 2018; Zrnic et al., 2021). In this work, we use the standard game-theoretic Stackelberg model proposed in Hardt et al. (2016) to simulate the agent's best response actions when choosing between recourse and manipulation. Our work considers the *imitation-based* manipulations: agents do not know the classifier f but are

---

[1]https://www.hud.gov/program-offices/housing/fhahistory

aware of a set of positively classified features and can misreport their feature by imitating another agent's feature that is positively classified. Such copycat behavior has been well-known in the literature of game theory, the behavioral economy, and strategic classification, e.g., (Bechavod et al., 2022; Barsotti et al., 2022). While most of this line of work focuses on agents being strategic and could potentially modify their features to get a favorable prediction outcome, our work focuses on when the system is being strategic and potentially withholds recourse to the agents.

Subsidy, Penalty, and Auditing Our work relates to interventions aimed at (dis)incentivizing strategic behaviors. Most relevant is Hu et al. (2019), who also studies strategic behavior using subsidies. Penalties for misreporting (Hardt et al., 2016; Blocki et al., 2013) offer another way to discourage manipulation, encouraging agents to pursue recourse instead. Both subsidies and penalties can be viewed as tools to shift the balance between the cost of recourse and manipulation — penalties raise the cost of manipulation, while subsidies lower the cost of recourse. Estornell et al. (2023a) explores auditing as an intervention to promote recourse, assuming universal recourse availability. The implementation of penalties requires verification power, such as in tax systems where crosschecking reported income deters misreporting. Subsidies, however, could be financed by third-party entities like governments or financial institutions. Incentivizing recourse through penalties is not ideal, as verification can lead to false positives, unfairly penalizing truthful agents. Audit-based systems (Estornell et al., 2023a) typically impose large fines, which can harm innocent agents if they are wrongly identified as manipulators. Subsidies avoid this issue. If the system controls audits and subsidies alone, it will prioritize its utility, which may not always align with the population's best interests.

3 PRELIMINARIES

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \equiv \{0,1\}$ be a domain of features and labels respectively. Let $f : \mathcal{X} \to \mathcal{Y}$ be a fixed binary classifier. A population of agents with features $\mathbf{X} = \{x : x \in \mathcal{X}\}$ and labels $Y = \{y : y \in \mathcal{Y}\}$ are classified by f, which is unknown to the agents; all agents desired to be positively classified (e.g., all loan applicants desire approval). Denote the domain of negatively classified features as $\mathcal{X}_- \subseteq \mathbb{R}^d$ and the domain of positively classified features as $\mathcal{X}_+ \in \mathbb{R}^d$, i.e. $f(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}_-$ and $f(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}_+$. All agents prefer positive classification over negative classification. Agents who have features $\mathbf{x} \in \mathcal{X}_-$ have two means of obtaining positive classification in the

next step, *recourse* and *manipulation*, which are defined next.

Recourse Recourse provides agents who received undesirable outcomes with recommended actions to genuinely improve their outcome by modifying their attributes (Ustun et al., 2019). Let $c_R : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be the cost of recourse, i.e. an agent with true features $\mathbf{x}$ pays cost $c_R(\mathbf{x}, \mathbf{x}')$ when modifying their features to be $\mathbf{x}'$. An agent with true feature $\mathbf{x} \in \mathcal{X}_-$ has an *optimal* recourse action [2],

$$\mathbf{x}_R(\mathbf{x}) = \operatorname{argmin}_{\mathbf{x}' \in \mathcal{X}_+} c_R(\mathbf{x}, \mathbf{x}') \qquad (1)$$
$$\text{s.t. } f(\mathbf{x}') = 1, \quad \mathbf{x}' \in A(\mathbf{x})$$

where $A(\mathbf{x})$ represents the set of features an agent with true features $\mathbf{x}$ can feasibly obtain, i.e., the *actionable* recourse actions provided by the system. When agents perform recourse, both their true features and true qualification rate change, i.e., their true features become $\mathbf{x}_R(\mathbf{x})$, and their true qualification rate changes from $\Pr(y = 1|\mathbf{x})$ to $\Pr(y = 1|\mathbf{x}_R(\mathbf{x}))$.

Manipulation In addition to recourse, agents can also perform manipulations. Following Barsotti et al. (2022), we focus on *imitation-based* manipulations: agents do not know the classifier f, but are aware of a set of publically revealed positively classified features $\mathbf{Z} \subseteq \mathcal{X}_+$ (defined below) and can misreport their feature by imitating another agent's feature that is positively classified and is publically revealed. For a manipulation cost function $c_M : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ the *optimal* imitation-based manipulation for an agent with true feature $\mathbf{x}$ is

$$\mathbf{x}_M(\mathbf{x}) = \arg \min_{\mathbf{x}' \in \mathbf{Z}} c_M(\mathbf{x}, \mathbf{x}') \qquad (2)$$

Different from recourse, manipulation is simply a misreport rather than a change of one's features, thus it does not change $\Pr[y = 1|\mathbf{x}]$. However, since the system only observes the reported features before classification, it does not know whether a report is truthful.

Feature Disclosure and Publicaly Revealed Set Z We model the set of publicly revealed features $\mathbf{Z} \subseteq \mathcal{X}_+$ resulting from agents sharing information with each other. In particular, $\mathbf{Z}$ consists of features that

---

[2]Throughout the paper, we will interchangeably use the terms 'recourse action' and 'recourse feature.' They both refer to the feature vector that will be classified positively after the agent's taking a particular recourse action. In other words, we assume that whenever an agent reveals their recourse *action*, it also reveals their original feature vector, which is equivalent to revealing the feature vector that corresponds to the vector *after* the agent performs recourse.
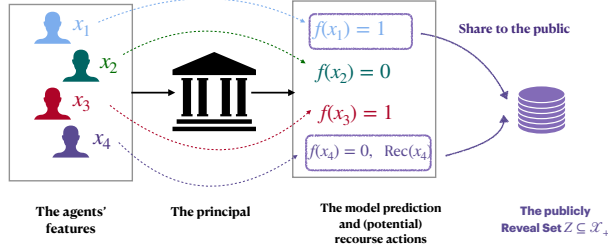
Figure 1: Demonstration of our modeling framework. Agents arrive simultaneously, and the system trains a classifier $f : \mathcal{X} \to \mathcal{Y}$ for maximum prediction accuracy. Negatively classified agents request recourse, and the system selects agents for recourse provision to maximize utility (Equation (3)). Positively classified agents and those provided recourse have a probability $p \in [0, 1]$ to reveal features, contributing to the publicly revealed set $\mathbf{Z} \subseteq \mathcal{X}_+$. Upon observing $\mathbf{Z}$, agents execute final actions based on Equation (5).

may come from two sets: 1) the revealed recourse actions recommended by the system (i.e., $\mathbf{z} \in \mathbf{X}_R$ where $\mathbf{X}_R = \{\mathbf{x}_R(\mathbf{x}), \mathbf{x} \in \mathbf{X}_-\}$), and 2) the set of initial positively classified features (i.e., $\mathbf{z} \in \mathbf{X}_+$). Each element is made public with a *fixed* probability $p \in [0, 1]$, and all publicly revealed elements make the reveal set $\mathbf{Z}$. We represent the set of recourse actions that are actually revealed as $\mathbf{Z}_R = \{\mathbf{z} \in \mathbf{X}_R : \text{Reveal}(\mathbf{z}) = 1\}$. Here, $\text{Reveal}(\mathbf{z})$ is a random indicator function that equals 1 with probability p (indicating that feature $\mathbf{z}$ is revealed) and 0 otherwise. Similarly, let $\mathbf{Z}_+$ represent the positively classified features that are actually revealed: $\mathbf{Z}_+ = \{\mathbf{z} \in \mathbf{X}_+ : \text{Reveal}(\mathbf{z}) = 1\}$. As a result, $\mathbf{Z} = \mathbf{Z}_R \cup \mathbf{Z}_+$.

This captures real-life scenarios where negatively classified agents collectively gather information about classifier f by observing positively classified peers or those who obtained recourse. Revealed recourse features are particularly crucial as they lie near the decision boundary, making them more likely targets for manipulation than general positive features.

4   INTERACTION BETWEEN AGENTS AND THE SYSTEM

Unlike the traditional recourse setting, where the system is expected to provide recourse to any individual upon request, without external regulation (e.g., government mandates requiring banks to offer recourse), a utility-maximizing system may have incentives to withhold recourse to prevent strategic manipulation by agents. In this section, we introduce our modeling framework to capture these dynamics.

A Motivating Example   *A bank publishes a classifier to determine who qualifies for a credit card. Each*

applicant (with feature vector $\mathbf{x}$) is approved if the bank's model f predicts they can repay their loan. For applicants denied a card, the bank may offer recourse, i.e., a plan to improve their creditworthiness, such as paying off debt or increasing their income. These recourse actions are provided through specific programs, such as financial classes. Agents also have access to an online forum where some applicants share their approved loan or recourse features. With knowledge of both recourse actions and the forum, some agents may misreport their features to match positively classified ones in an attempt to gain approval without actually taking the recommended recourse actions. As a result, the bank may have an incentive to limit recourse to individuals whose features are harder to manipulate (e.g., features that are easier for the bank to verify).

We now formalize the dynamics between the recourse system and the agents.

System: The system trains a classifier $f : \mathcal{X} \to \mathcal{Y}$ to maximize the prediction accuracy:

$$f = \arg\max_{f \in \mathscr{F}} \sum_{x \in \mathbf{X}} \mathbb{1}[f(\mathbf{x}) = y]$$

A collection of negatively classified agents with features $\mathbf{X}_- \subseteq \mathcal{X}_-$ will request recourse actions from the system after receiving their prediction outcome. The system first computes optimal recourse actions for all negatively classified agents but only chooses to *release* a subset of those recourse actions $\mathbf{Z}_R \subset \mathbf{X}_R$ to the public to maximize its utility, i.e., TP − FP:

$$\max_{\mathbf{Z}_R \subset \mathbf{X}_R} \underbrace{\mathsf{TP}(\mathbf{S}) - \mathsf{FP}(\mathbf{S})}_{\text{system's utility}} \tag{3}$$

$$\text{s.t.} \quad \underbrace{\mathbf{S} = \{\mathbf{z}(\mathbf{x}, \mathbf{Z}) : \mathbf{x} \in \mathbf{X}\}}_{\text{agent's reported features (Eq 5)}} \tag{4}$$

$$\underbrace{\mathbf{Z} = \mathbf{Z}_R \cup \mathbf{Z}_+}_{\text{all publicly revealed features}}$$

Here, $\mathsf{TP}(\mathbf{S})$ and $\mathsf{FP}(\mathbf{S})$ are the true positive and false positive rates on the set of features after the agent's final actions. We assume that the system either knows c_R and c_M, or can reasonably approximate these cost functions when optimizing their objective. Intuitively, this definition of system utility reflects a bank gaining a utility of 1 for each repaid loan and −1 for each defaulted loan.

Agents: Agents who are negatively classified will request a recourse action from the system. Upon seeing the publically revealed features $\mathbf{Z}$ defined in Section 3, agents who are provided with a recourse action adapt their features from $\mathbf{x}$ to $\mathbf{z} = \mathbf{x}_M(\mathbf{x})$ or $\mathbf{z} = \mathbf{x}_R(\mathbf{x})$ such that $f(\mathbf{z}) = 1$, while minimizing the cost of the

corresponding action. When both the recourse and manipulation actions are greater than 1[3], the agents will choose to stay with their original features $\mathbf{x}$, which corresponds to the *do-nothing* action. Agents who are not provided with a recourse action will choose to manipulate or *do nothing*. The final action for already positively classified agents is always the *do-nothing* action.

Agent's best response: Denote $\zeta_{\mathbf{x}} \in \{0, 1\}$ as an indicator for whether agent $\mathbf{x}$ is provided with a recourse or not (i.e., $\zeta(\mathbf{x}) = 1$ when provided with a recourse action). Then for all agents with $f(\mathbf{x}) = 0$, their final action is:

$$\mathbf{z}(\mathbf{x}, \mathbf{z}) = \begin{cases} \mathbf{x}_R(\mathbf{x}) & \zeta_{\mathbf{x}} = 1 \text{ and } c_R(\mathbf{x}, \mathbf{x}_R(\mathbf{x})) < \min(1, c_M(\mathbf{x}, \mathbf{x}')), \\ & \forall \mathbf{x}' \in \mathbf{Z} \\ \mathbf{x}_M(\mathbf{x}) & \zeta_{\mathbf{x}} = 1 \text{ and } c_M(\mathbf{x}, \mathbf{x}_M(\mathbf{x})) < \min(1, c_R(\mathbf{x}, \mathbf{x}')), \\ & \forall \mathbf{x}' \in \mathbf{Z}, \text{ or } \zeta_{\mathbf{x}} = 0 \text{ and } c_M(\mathbf{x}, \mathbf{x}_M(\mathbf{x})) < 1 \\ \mathbf{x} & \zeta_{\mathbf{x}} = 1 \text{ and } c_R(\mathbf{x}, \mathbf{x}_R(\mathbf{x})), c_M(\mathbf{x}, \mathbf{x}_R(\mathbf{x})) \geq 1, \\ & \forall \mathbf{x}' \in \mathbf{Z}, \text{ or } \zeta_{\mathbf{x}} = 0 \text{ and } c_M(\mathbf{x}, \mathbf{x}_M(\mathbf{x})) \geq 1 \end{cases}$$
(5)

Summary of System-Agent Interaction:

1. Agents arrive simultaneously, and the system trains a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ for maximum prediction accuracy.

2. Negatively classified agents request recourse, and the system selects agents for recourse provision to maximize utility (Equation (3)).

3. Positively classified agents and those provided recourse have a probability $p \in [0, 1]$ to reveal features, contributing to the publicly revealed set $\mathbf{Z} \subseteq \mathcal{X}_+$.

4. Upon observing $\mathbf{Z}$, agents execute final actions based on Equation (5).

Our framework is intended to capture settings where black box models are used for decision-making. Any agent subjected to the decision rules will not have direct access to the model but will still act in their best interest. In these opaque settings, recourse proposed by the system naturally offers a way for agents to learn more about the decision rule, thus increasing their ability to game the system.

The following two definitions introduce key metrics that will be used throughout this paper – the recourse rate quantifies the proportion of negatively classified

---

[3]The strategic agent's utility for adapting their feature from x to x' is determined by the standard utility function in the literature of strategic classification (see, e.g., Hardt et al. (2016)), which is $U(x, x') = f(x') - c(x, x')$. Thus, when the cost of adaptation $c(x, x') \geq 1$, the utility will be less than 0, in which case, the agent does nothing.

agents who opt to take recourse actions when presented with a disclosed feature set. The manipulation rate captures the fraction of negatively classified agents who choose to manipulate their features under the same conditions:

Definition 1 (*Recourse Rate*) *Let* $\mathbf{X}_-$ *be the set of features of negatively classified agents. For a given set of disclosed features (i.e., recourse actions)* $\mathbf{Z}$*, the* recourse rate $rec(\mathbf{Z}, \mathbf{X}_-)$ *is defined as the fraction of agents who choose to perform recourse when shown* $\mathbf{Z}$*:*

$$rec(\mathbf{Z}, \mathbf{X}_-) = \frac{\sum\limits_{\mathbf{x} \in \mathbf{X}_-} \mathbb{1}\left[\min\limits_{\mathbf{z}' \in Z} c_R(\mathbf{x}, \mathbf{z}') < \min\left(1, \min\limits_{\mathbf{z}'' \in Z} c_M(\mathbf{x}, \mathbf{z}'')\right)\right]}{|\mathbf{X}_-|}$$

Definition 2 (*Manipulation Rate*) *Let* $\mathbf{X}_-$ *be the set of features of the negatively classified agents. For a given set of disclosed features (i.e., recourse actions)* $\mathbf{Z}$*, the* manipulation rate $manip(\mathbf{Z}, \mathbf{X}_-)$ *is defined as the fraction of the n agents which choose to manipulate when shown features* $\mathbf{Z}$*:*

$$manip(\mathbf{Z}, \mathbf{X}_-) = \frac{\sum\limits_{\mathbf{x} \in \mathbf{X}_-} \mathbb{1}\left[\min\limits_{\mathbf{z}' \in Z} c_M(\mathbf{x}, \mathbf{z}') < \min\left(1, \min\limits_{\mathbf{z}'' \in Z} c_R(\mathbf{x}, \mathbf{z}'')\right)\right]}{|\mathbf{X}_-|}$$

5 SYSTEM UTILITY

Recall from the previous section, the system aims to select a set $\mathbf{Z}_R \subseteq \mathbf{X}_R$ to reveal as recourse recommendations simultaneously to maximize its utility (Equation (3)). We can first show that this problem is NP-hard (Theorem 6 in Appendix C.2). Despite the hardness of this objective, the system's utility is *submodular* in the set of provided recourse actions (Theorem 7 in Appendix C.3). This characteristic enables the system to employ standard submodular optimization techniques to approximately get the optimal recourse actions to disclose to k agents.

We can show that in expectation, the system benefits from agents *taking* recourse actions:

Theorem 1 (*System's Expected Utility Changes*) *The system's expected utility (defined in Eq. (3)) increases for each recourse action* taken *by agents and decreases for every manipulation action* taken *by agents. When the classifier used by the system is better than random guessing, which means that $f(x) = 1$ implies $\Pr[y(x) = 1 | X = x] \geq 0.5$, then the system's utility is monotonically increasing in each recourse action taken by an agent in expectation but will be monotonically decreasing in each manipulation action taken by an agent.*

However, this does not imply that the system is always incentivized to provide as many recourse actions as possible, since agents might not always take them

if they collude, which creates a natural misalignment between the system's utility and recourse offering for the system.

6 COST OF STRATEGICALLY WITHHOLDING RECOURSE SYSTEM

Having shown that the system could be incentivized to withhold recourse from the agents, we now study the consequence of such withholdings by examining the social cost and unfairness as a result of the system's strategic actions.

Definition 3 *(Social Cost of a Strategically Withhold Recourse) Given a publically revealed set $\mathbf{Z} \subseteq \mathcal{X}_+$, the* social cost *refers to the additional cost agents must pay as a result of the system withholding recourse. Denote $\mathbf{x}_R(\mathbf{x})$ as the optimal recourse action provided by a non-strategic system, and $\mathbf{z}_R(\mathbf{x}, \mathbf{Z})$ as the recourse action that the agent takes given the revealed set $\mathbf{Z}$, then the social cost of a strategically withholding recourse system is defined as:*

$$cost(\mathbf{Z}, \mathbf{X}_-) = \sum_{\mathbf{x} \in \mathbf{X}_-} \left(c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}, \mathbf{Z})) - c_R(\mathbf{x}, \mathbf{x}_R(\mathbf{x})) \right)$$

where $\mathbf{z}_R(\mathbf{x}, \mathbf{Z}) = \arg\min_{\mathbf{z} \in \mathbf{Z}} c_R(\mathbf{x}, \mathbf{z})$. For the remainder of our results, we focus on univariate classifiers, i.e., the feature $\mathbf{x}$ is one-dimensional. There is a natural correspondence between univariate and multivariate classifiers in the sense that one can imagine the space of single-dimensional features as the scores produced a multi-dimensional classifier $f(\mathbf{x})$ [4]. That is, in the case when $f(\mathbf{x}) = [h(\mathbf{x}) \geq \theta]$ for some score function h and threshold theta, we can view f as a single dimensional classifier acting on the space of scores produced by h.

We also measure the disparities of different social groups in terms of their differences in 1) recourse ratios (defined in Definition 1), and 2) social cost (defined in Definition 3). Understanding the disparities in terms of recourse rate and social cost among different groups is crucial for addressing issues of unfairness in an algorithmic recourse system Gupta et al. (2019); von Kügelgen et al. (2022). These disparities often reflect systemic biases and inequalities, impacting marginalized communities disproportionately. In particular, assume there are two groups of agents $\mathbf{X}^{(g_0)}$ and $\mathbf{X}^{(g_1)}$, where g_0, g_1 represents their group memberships, we are interested in the following quantities:

---

[4]This follows similarly to Lemma 3.1 in Milli et al. (2019).

Definition 4 *(Disparity in Social Cost and Recourse Ratio) The disparity in social cost and recourse ratio for two groups g_0, g_1 are defined as:*

$$Diff^{(cost)}(\mathbf{Z}, \mathbf{X}^{(g_0)}, \mathbf{X}^{(g_1)}) := \left| cost(\mathbf{Z}, \mathbf{X}_-^{(g_1)}) - cost(\mathbf{Z}, \mathbf{X}_-^{(g_0)}) \right|,$$

$$Diff^{(rec)}(\mathbf{Z}, \mathbf{X}^{(g_0)}, \mathbf{X}^{(g_1)}) := \left| rec(\mathbf{Z}, \mathbf{X}_-^{(g_1)}) - rec(\mathbf{Z}, \mathbf{X}_-^{(g_0)}) \right|$$

In the experiments section, we demonstrate that these disparities can be quite common across different datasets (see Figure 3). By quantifying and illuminating these disparities, we gain crucial insights into the specific mechanisms of inequity and injustice within algorithmic recourse systems.

7 THE EFFECT OF SUBSIDIES

To remedy the adverse population- and group-level impacts previously observed, we investigate the use of subsidies (rigorously defined next) and their impact on recourse rate, social cost, and unfairness we defined in the previous section. Subsidies correspond to a global decrease in the cost of recourse. For example, free educational material on financial literacy distributed to any agent petitioning the bank for recourse will increase the ease at which that agent can perform recourse actions.

Definition 5 *(Subsidies)(Hu et al., 2019) A subsidy $0 \leq \alpha \leq 1$ is a scalar decrease to the cost of recourse. For subsidy α, agents performing recourse pay only $(1 - \alpha) \cdot c_R(\mathbf{x}, \mathbf{x}')$ instead of the full cost of $c_R(\mathbf{x}, \mathbf{x}')$. We denote $c_R(\mathbf{x}, \mathbf{x}'; \alpha) = (1 - \alpha) \cdot c_R(\mathbf{x}, \mathbf{x}')$ as the new recourse cost at subsidy level α.*

Next, we demonstrate how subsidies can help increase the recourse rate (Theorem 1) and system's utility (Theorem 3). Additionally, subsidies can mitigate disparities in recourse rate differences (Theorem 5) and social cost differences (Theorem 4) among various groups.

We first show how subsidies influence the recourse rate. Recall that subsidy reduces the cost of recourse from $c_R(\mathbf{x}, \mathbf{x}')$ to $c_R(\mathbf{x}, \mathbf{x}'; \alpha)$. With that, the recourse rate becomes:

$$rec(\mathbf{Z}, \mathbf{X}_-; \alpha)$$

$$= \frac{\sum_{\mathbf{x} \in \mathbf{X}_-} \mathbb{1}\left[\min_{\mathbf{z}' \in \mathbf{Z}} c_R(\mathbf{x}, \mathbf{z}'; \alpha) < \min\left(1, \min_{\mathbf{z}'' \in \mathbf{Z}} c_M(\mathbf{x}, \mathbf{z}'')\right) \right]}{|\mathbf{X}_-|}.$$

The key observation here is that with subsidy α, the recourse cost reduces, but the manipulation cost remains the same. Both optimal recourse actions $\mathbf{x}_R(\mathbf{x})$ and the optimal manipulation action $\mathbf{x}_M(\mathbf{x})$ remain

the same. With that, we can show that the recourse rate is a monotonic function in subsidy – as the subsidy level increases, the recourse rate will also increase:

Theorem 1 *(Subsidy Influence on Recourse Rate) Given a reveal set $\mathbf{Z}$, the recourse rate $rec(\mathbf{Z}, \mathbf{X}_-, \alpha)$ is a monotonically increasing function of subsidies α.*

With subsidy α, the social cost for a given revealed set $\mathbf{Z}$ becomes:

$$\text{cost}(\mathbf{Z}, \mathbf{X}_-; \alpha) = \sum_{\mathbf{x} \in \mathbf{X}_-} \left(c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}, \mathbf{Z}; \alpha); \alpha) - c_R(\mathbf{x}, \mathbf{x}_R; \alpha) \right)$$

where $\mathbf{z}_R(\mathbf{x}, \mathbf{Z}; \alpha) = \arg \min_{\mathbf{z} \in \mathbf{Z}} (1 - \alpha) c_R(\mathbf{x}, \mathbf{z})$ is the optimal recourse action given revealed set $\mathbf{Z}$ and a particular subsidy level α, and $\mathbf{x}_R$ is the optimal default recourse action provided by the system without any strategic withholding. We can show that the social cost is also a monotonic non-increasing function in the subsidy level:

Theorem 2 *(Subsidy Influence on Social Cost) Given a revealed set $\mathbf{Z}$, the social cost $cost(\mathbf{Z}, \mathbf{X}_-; \alpha)$ is monotonically decreasing in subsidies.*

Subsidies also help improve the system's utility; under some assumptions on the cost functions (i.e., monotonic in the distance and only cross once), the system's utility is monotonic in subsidies as well:

Theorem 3 *(Subsidy's Influence on System's Utility) Given a revealed set $\mathbf{Z}$, when both $c_R(\mathbf{x}, \mathbf{x}')$ and $c_M(\mathbf{x}, \mathbf{x}')$ are monotonic in $\|\mathbf{x} - \mathbf{x}'\|$ and only cross once, the system utility is monotonically increasing in subsidies.*

Next we examine the difference in social cost between groups as a function of subsidies. We find that subsidies are an effective tool to mitigate disparities caused by strategically withheld recourse.

Theorem 4 *(Subsidy Influence on Social Cost Disparity) With subsidy α, the disparity in social cost for two group g_0, g_1 becomes: $Diff^{(cost)}(\mathbf{Z}, \mathbf{X}^{(g_0)}, \mathbf{X}^{(g_1)}; \alpha) := \left| cost(\mathbf{Z}, \mathbf{X}_-^{(g_1)}; \alpha) - cost(\mathbf{Z}, \mathbf{X}_-^{(g_0)}; \alpha) \right|$ Given a revealed set $\mathbf{Z}$, the social cost difference monotonically decreases in subsidies.*

Intuitively, as we increase the subsidy level, the cost of recourse decreases linearly as a function of the subsidy level, making it increasingly cheaper to perform the optimal recourse action. For both social groups, their social cost approaches 0 as we increase the subsidy level; as a result, the disparity in social cost between the two groups also decreases to 0.

With subsidy α, for a given a revealed set $\mathbf{Z}$, the disparity in recourse ratio for groups g_0, g_1 is:

$$Diff^{(\text{rec})}(\mathbf{Z}, \mathbf{X}^{(g_0)}, \mathbf{X}^{(g_1)}; \alpha) := \left| \text{rec}(\mathbf{Z}, \mathbf{X}_-^{(g_1)}; \alpha) - \text{rec}(\mathbf{Z}, \mathbf{X}_-^{(g_0)}; \alpha) \right|$$

where $\text{rec}(\mathbf{Z}, \mathbf{X}_-^{(g_i)})$ is the recourse rate for a particular subgroup g_i. We show that when subsidies are sufficiently large, the recourse rate difference is monotonically decreasing in subsidies:

Theorem 5 *(Subsidy's Influence on Recourse Rate Disparity) Given two groups g_0 and g_1 of relatively equal negatively classified agents size $|\mathbf{X}_-^{(g_0)}| \approx |\mathbf{X}_-^{(g_1)}|$, there exists a subsidy level $0 \le \alpha^* \le 1$, such that $\forall \alpha \ge \alpha^*$, the recourse rate difference monotonically decreases.*

This result follows that when recourse is free, i.e., subsidies are maximized, all agents can perform recourse, and the recourse rate difference is 0. Thus, as subsidies increase, there must exist a point (namely α^*) when both groups can take advantage of subsidies at proportional rates, thus decreasing the gap between the number of agents performing recourse in both groups. We also verify empirically that for recourse rate difference, there indeed exists a peak subsidy value α^* where the recourse rate difference increases before and then decreases afterward (see Figure 4).

8 EMPIRICAL STUDIES

Setup We conduct experiments using three datasets: 1) **Law School** Wightman & Council (1998) dataset, in which the objective is to predict whether a student will pass the bar exam on the first attempt, **Adult Income** Dua et al. (2017) in which the objective is to predict whether an individual earns more than $50K$ annually, and **German Credit** Yeh & Lien (2009) in which the objective is to predict whether a given individual will *not* default on their credit. In each dataset, agents have constant utility over approved features, i.e., the conventional recourse setting where $u_a(\mathbf{x}) = 1$ for all $\mathbf{x}$; the principal (system) has utility $u_p(\mathbf{x}) = 1$ when the agent is a true positive ($y = 1$, $f(x) = 1$) and $u_p(\mathbf{x}) = -1$ when the agent is a false positive ($y = -1$, $f(x) = 1$). Qualification is predicted via Logistic Regression (shown in this section) or Gradient Boosting Trees (shown in the Supplement Appendix F).

Recourse and manipulation both carry an ℓ_2 cost, namely $c_R(\mathbf{x}, \mathbf{z}) = \|w_R \cdot (\mathbf{x} - \mathbf{z})\|_2$, and $c_M(\mathbf{x}, \mathbf{z}) = \|w_M \cdot (\mathbf{x} - \mathbf{z})\|_2$, where w_R and w_M are the weight vectors for the cost functions. In our experiments, we report outcomes over 100 runs using randomly initialized w_R and w_M and resampled subsets of positive and negative agents in the dataset in each run. We set the probability that the agent discloses their feature publicly at $p = 0.7$ for all experiments. When varying this value, we observe similar results.
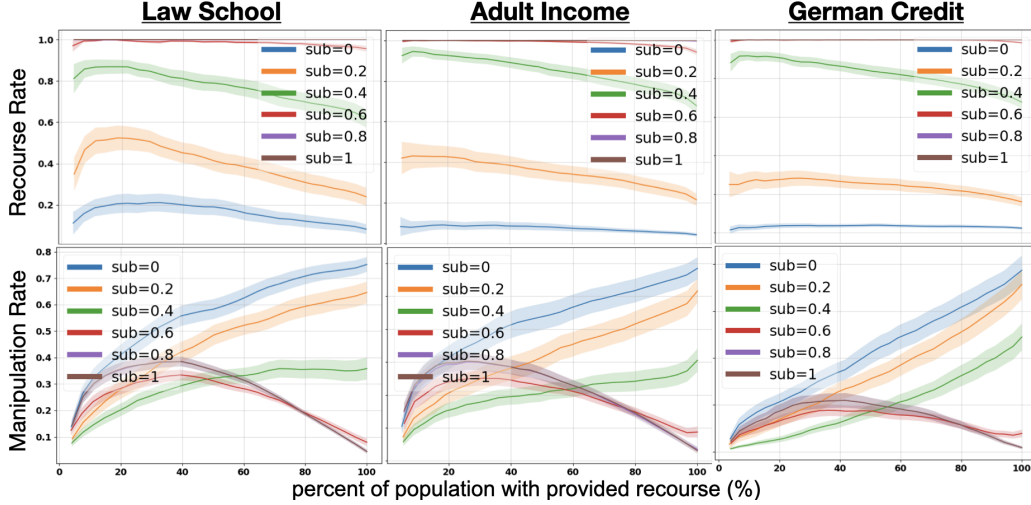
Figure 2: *Fraction of the population performing recourse (top row) or manipulation (bottom row). Each line corresponds to a different subsidy ratio "sub", i.e., the cost reduction applied to recourse.*
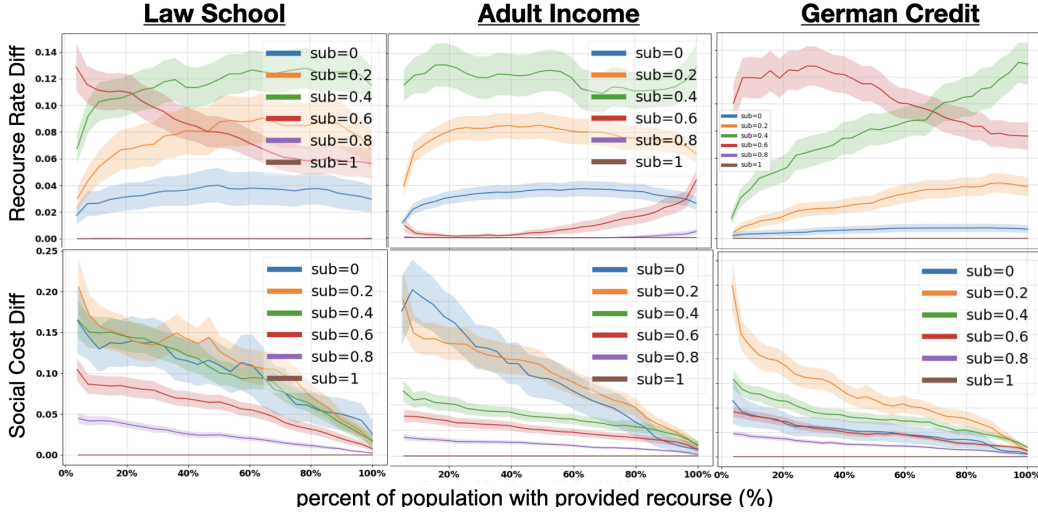

Figure 3: *Difference in recourse rate (top row) and social cost (bottom row) between different sensitive attribute groups. Each line corresponds to a different subsidy ratio "subs", i.e., the cost reduction applied to recourse.*

Recourse Rate and Manipulation Rate We begin by examining the relationship between the fraction of the population choosing to perform recourse and the fraction choosing to perform manipulation as a function of the fraction of agents given a recourse action. In Figure 2, we see that in general, as the percentage of agents who are provided a recourse action increases, the recourse rate decreases while the manipulation rate increases (this trend holds for each subsidy value). Thus, when agents themselves can strategically select between recourse and manipulation, the increased model transparency, created by providing more agents with recourse actions, results in more agents selecting to perform manipulation. Providing more recourse actions to agents, does not necessarily result in more agents performing recourse. Despite

this general trend, we also observe the effectiveness of subsidies. As subsidies converge to 1 (meaning recourse carries no cost), the fraction of agents choosing recourse converges to 1, while the fraction of agents choosing manipulation converges to 0. While it may be expensive in general to provide such subsidies, and the question of how to balance this expense against the system's own utility remains open, these results indicate that subsidies are an effective avenue for broadly promoting recourse and disincentivizing manipulation.

Disparity in Recourse and Social Cost Lastly, we investigate how strategic system behavior causes disparate impacts among sensitive groups. In our experiments, groups are taken to be binary and are defined by race in the Law School dataset (White and

Non-White), by gender in the Adult Income dataset (Male and Female), and by age in the German Credit dataset (Young and Old). In Figure 3, we see the difference in the number of agents performing recourse and social cost between groups. Higher values in these plots indicate higher rates of recourse, or lower cost, for White individuals in the Law School dataset, Male individuals in the Adult income dataset, and Young individuals in the Credit dataset. First, strong subsidies (particularly subs ≤ 0.4) result in a large decrease in the disparities between groups for both recourse rate and social cost. For less strong subsidies (subs ≥ 0.6), we see that the gap in recourse rate between groups can increase. This is due to the fact that when subsidies are less strong, only agents with already low costs of recourse (primarily from the advantaged group) can benefit from those subsidies.

9 CONCLUSION

In scenarios where agents can manipulate a system, there is a reduced incentive for the system to provide recourse due to increased model transparency. Consequently, the system strategically withholds recourse from some, leading to higher social costs, and disproportionately impacting disadvantaged groups. Despite the inherent tension between the system's utility and its provision of recourse, subsidies emerge as a viable tool to boost recourse-providing rates and alleviate group-wise disparities resulting from recourse withholding.

References

Emrah Akyol, Cedric Langbort, and Tamer Basar. Price of transparency in strategic machine learning. *arXiv preprint arXiv:1610.08210*, 2016.

Flavia Barsotti, Rüya Gökhan Koçer, and Fernando P Santos. Transparency, detection and imitation in strategic classification. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI 2022*. International Joint Conferences on Artificial Intelligence (IJCAI), 2022.

Yahav Bechavod, Chara Podimata, Steven Wu, and Juba Ziani. Information discrepancy in strategic learning. In *International Conference on Machine Learning*, pp. 1691–1715. PMLR, 2022.

Jeremiah Blocki, Nicolas Christin, Anupam Datta, Ariel D Procaccia, and Arunesh Sinha. Audit games. *arXiv preprint arXiv:1303.0356*, 2013.

Yatong Chen, Jialu Wang, and Yang Liu. Linear classifiers that encourage constructive adaptation. *arXiv preprint arXiv:2011.00355*, 2020.

Yatong Chen, Zeyu Tang, Kun Zhang, and Yang Liu. Model transferability with responsive decision subjects. In *International Conference on Machine Learning*, pp. 4921–4952. PMLR, 2023.

Yiling Chen, Chara Podimata, Ariel D Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 9–26, 2018.

Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70, 2018.

Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017.

Ahmad-Reza Ehyaei, Amir-Hossein Karimi, Bernhard Schölkopf, and Setareh Maghsudi. Robustness implies fairness in causal algorithmic recourse. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 984–1001, 2023.

Andrew Estornell, Yatong Chen, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Incentivizing recourse through auditing in strategic classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 400–408, 08 2023a. doi: 10.24963/ijcai.2023/45.

Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. Group-fair classification with strategic agents. In *ACM Conference on Fairness, Accountability, and Transparency*, pp. 389–399, 2023b.

Hidde Fokkema, Damien Garreau, and Tim van Erven. The risks of recourse in binary classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 550–558. PMLR, 2024.

Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

Keegan Harris, Valerie Chen, Joon Kim, Ameet Talwalkar, Hoda Heidari, and Steven Z Wu. Bayesian persuasion for algorithmic recourse. *Advances in Neural Information Processing Systems*, 35:11131–11144, 2022.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 259–268, 2019.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects, 2020a.

Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach, 2020b.

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

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 230–239, 2019.

Matthew Olckers and Toby Walsh. Incentives to offer algorithmic recourse, 2023.

Andrew Orso, Jon Lee, and Siqian Shen. Submodular minimization in the context of modern lp and milp methods and solvers. In *Proceedings of the 14th International Symposium on Experimental Algorithms - Volume 9125*, pp. 193–204, Berlin, Heidelberg, 2015. Springer-Verlag. ISBN 9783319200859.

Stratis Tsirtsis, Behzad Tabibian, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Optimal Decision Making Under Strategic Behavior. *arXiv e-prints*, 2019.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 10–19, 2019.

Suresh Venkatasubramanian and Mark Alfano. The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 284–293, 2020.

Julius von Kügelgen, Umang Bhatt, Amir-Hossein Karimi, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse, 2020.

Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse, 2022.

Yevgeniy Vorobeychik. The many faces of adversarial machine learning. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 15402–15409, 2023.

L.F. Wightman and Law School Admission Council. *LSAC National Longitudinal Bar Passage Study.* LSAC research report series. Law School Admission Council, 1998. URL https://books.google.com/books?id=O9A7AQAAIAAJ.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

Tijana Zrnic, Eric Mazumdar, Shankar Sastry, and Michael Jordan. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems*, 34:15257–15269, 2021.

To Give or Not to Give? The Impacts of Strategically Withheld Recourse

A Notation Table

Symbol	Usage
$\mathcal{X} \subset \mathbb{R}^d$	The domain of the feature $\mathbf{x}$
$\mathcal{Y} \equiv \{0,1\}$	The domain of labels
$\mathbf{X} \in \mathbb{R}^{n \times d}$	A set of features of n agents
$Y \in \{0,1\}^{|\mathbf{X}|}$	The labels for the set of features $\mathbf{X}$
$\mathbf{x} \in \mathcal{X}$	A random variable representing an example's features
$y \in \mathcal{Y}$	A random variable representing an example's *ground truth label*
$f : \mathcal{X} \to \mathcal{Y}$	a binary classifier, unknown to the agents
$\mathcal{X}_-, \mathcal{X}^{(0)} \subseteq \mathcal{X}$	The domain of negatively classified features, i.e. $\forall \mathbf{x} \in \mathcal{X}_-, f(\mathbf{x}) = 0$
$\mathcal{X}_+ \subseteq \mathcal{X}$	The domain of positively classified features, i.e., $\forall \mathbf{x} \in \mathcal{X}_+, f(\mathbf{x}) = 1$
$\mathbf{X}_- \subseteq \mathcal{X}$	The set of negatively classified features, i.e. $\forall \mathbf{x} \in \mathbf{X}_-, f(\mathbf{x}) = 0$
$\mathbf{X}_+ \subseteq \mathcal{X}$	The set of positively classified features, i.e., $\forall \mathbf{x} \in \mathbf{X}_+, f(\mathbf{x}) = 1$
$\mathbf{X}^{(g_i)} \subseteq \mathbf{X}$	The subset of features belongs to group $G = g_i$
$c_R : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$	The cost function of recourse
$c_M : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$	The cost function of recourse
$\mathbf{X}_R$	The set of all possible recourse actions
$\mathbf{Z}_R$	The set of revealed recourse actions
$\mathbf{Z}_+$	The set of revealed positively classified features
$\mathbf{Z} = \mathbf{Z}_R \cup \mathbf{Z}_+$	A publicly revealed feature set
$\mathbf{x}_R(\mathbf{x})$	The optimal recourse action for agent with feature $\mathbf{x}$
$\mathbf{x}_M(\mathbf{x})$	The optimal manipulation action for agent with feature $\mathbf{x}$
$\mathbf{z}(\mathbf{x}, \mathbf{Z})$	The agent's final action
$\mathrm{rec}(\mathbf{Z}, \mathbf{X})$	The recourse ratio for feature sets $\mathbf{X}$ given revealed set is $\mathbf{Z}$
$\alpha \in [0,1]$	A subsidy level
$u_0 \in \mathbb{R}$	The initial utility of a system without providing recourse

Table 1: Primary Notation

B   Additional Related Work

Recourse   Recourse focuses on providing agents with the ability to contest or improve their outcome via a modification to their attributes in a *genuine* manner (e.g., paying off debt to increase creditworthiness) Ustun et al. (2019); Venkatasubramanian & Alfano (2020); Karimi et al. (2020a); Gupta et al. (2019); Karimi et al. (2020b); von Kügelgen et al. (2020); Chen et al. (2020); Harris et al. (2022). Much of this line of work focuses on the setting where the requested recourse is guaranteed to be provided. As far as we know, our work is the first to challenge this fundamental assumption and argue that without a third-party's intervention (e.g., the government regulation on the system's recourse providing), a utility-maximizing algorithmic recourse system may be incentivized to strategically withhold recourse from some agents to prevent manipulations. We point the reader to Karimi et al. (2020a) for a more detailed discussion of the concepts and recent development of algorithmic recourse. To our knowledge, even though the literature has previously introduced the concepts of recourse, strategic manipulation, and subsidy analysis, we have yet to find any studies that explicitly highlight how a recourse system might act strategically by withholding recourse to enhance its own utility. The originality of our work, thus, is to address this gap. Some works have investigated the relationship between incentives/utility and recourse, such as Fokkema et al. (2024), which finds that providing recourse can decrease classifier accuracy, Estornell et al. (2023a), which investigates ways to ensure that given recourse actions are taken by agents, and Olckers & Walsh (2023) which investigates the incentive compatibility of recourse.

Strategic Classification   Strategic Classification focuses on the problem of how to effectively make predictions in the presence of agents who behave strategically to obtain desirable outcomes Hardt et al. (2016); Chen et al. (2018); Tsirtsis et al. (2019); Levanon & Rosenfeld (2021); Dong et al. (2018); Chen et al. (2018); Zrnic et al. (2021); Chen et al. (2023). Our work considers a specific type of strategic behavior, namely the *imitation-based* manipulations: agents do not know the classifier f but are aware of a set of positively classified features and can misreport their feature by imitating another agent's feature that is positively classified. Such copycat behavior has been well-known in the literature of game theory, the behavioral economy, and strategic classification, e.g., Bechavod et al. (2022); Barsotti et al. (2022). While most of this line of work focuses on agents being strategic and could potentially modify their features to get a favorable prediction outcome, our work focuses on when the system is being strategic and potentially withholds recourse to the agents.

Fairness and Social Cost in Recourse and Strategic Classification   Fairness has been explored in the literature algorithmic recourse and strategic classification. For example, existing works on fairness in recourse emphasize the importance of equitable recourse and explore various remedying unfair recourse decisions Gupta et al. (2019); von Kügelgen et al. (2022); Ehyaei et al. (2023). Fairness with the presence of strategic behavior has featured studies that highlight the inequity that results from strategic behavior by individuals Hu et al. (2019), as well as inequity (e.g., social cost) resulting from making classifiers robust to strategic behavior Milli et al. (2019); Estornell et al. (2023b). Unlike previous work that primarily focuses on proposing fair classifiers with the presence of strategic agents, our work uniquely demonstrates how the system's strategic withholding impacts the fairness and social cost for different societal groups.

Transparency   Also related is work on transparency in machine learning. In particular, Barsotti et al. (2022) find that the risks of transparent explanations are alleviated if effective methods to detect faking behaviors are in place. Unlike our modeling framework, they model transparency as how much noise is in the threshold of a threshold classifier. Akyol et al. (2016) examines the impact of users' strategic behavior on the design and performance of transparent machine learning algorithms, quantifying the "price of transparency" as the cost ratio for the algorithm designer when users exploit transparency compared to when the algorithm is opaque.

Comparison with three closely related papers

- Comparison with Estornell et al. (2023a): the key distinction between these our work and Estornell et al. (2023a) is that Estornell et al. (2023a) presumes the system will provide any agent with an optimal recourse action and examines how auditing can dissuade agents from manipulating. In contrast, we do not consider auditing, and instead focus on how a system may be incentivized to withhold recourse from certain agents. While both papers examine the use of subsidies, our focus and model are distinct.

- Comparison with Fokkema et al. (2024): Fokkema et al. (2024) focuses on discussing the accuracy drop

as a result of the system providing recourse, because it pushes users to regions of higher class uncertainty and therefore leads to more mistakes. Our work, on the other hand, focuses on the incentive-compatibility problem in an algorithmic recourse system. of the population

- Comparison with Olckers & Walsh (2023): similar to our work, Olckers & Walsh (2023) also studies when it is incentive-compatible for a decision-maker to offer recourse. Unlike our setting, however, they primarily operate on a simple toy model that assumes the applicant's *profitability* is fixed.

C   Proofs for Section 5

C.1   ILP for system when p = 1

We provide the ILP formula for the system to find optimal recourse actions when the revealing probability $p = 1$:

$$\max_{\mathbf{a} \in \{0,1\}^{|\mathbf{Z}_{\max}|}, \mathbf{b} \in \{0,1\}^{|\mathbf{X}_-|}} \sum_{j=1}^{|\mathbf{X}_-|} b_j \qquad \text{(maximize the number of agents performing recourse)}$$

$$\text{s.t.} \quad b_j c_R(\mathbf{x}_j, \mathbf{z}_R) \leq a_i c_M(\mathbf{x}_j, \mathbf{z}_i) + (1 - a_i)$$
$$\text{(only do recourse if all manipulation costs are greater)}$$

$$b_j \leq a_{j_R} \qquad \text{(the optimal recourse action } \mathbf{z}_{j_R} \text{ for agent } j \text{ must be revealed)}$$

$$b_j c_R(x_j, z_R) \leq 1 \qquad \text{(the optimal recourse action } \mathbf{z}_{j_R} \text{ for agent } j \text{ must be less than 1)}$$

$$\sum_{h=1}^{|\mathbf{Z}|} a_h = k \qquad \text{(the total number of revealed recourse action is } k)$$

C.2   NP hardness of the System's Optimal Recourse Providing Problem

Theorem 6 *The problem of selecting the optimal set of recourse actions to recommend, such that the system's utility is maximized (Equation 3), is NP-hard, even when the probability of diclosure $p = 1$.*

Proof 1 *To demonstrate the intractability of this objective, we reduce from the known NP-hard problem Minimum k-Union (MkU), an instance of which is defined via a universe of n elements $U = \{s_1, \ldots, s_n\}$, a collection of n sets $\mathbf{S} = \{S_1, \ldots S_m\}$ with elements in U, and a budget k. The objective in MkU is to select an index set I of size exactly k such that $\cup_{j \in I} S_j$ is minimized. Given an instance of MkU can be mapped to an instance of simultaneous recourse as follows. Let $\mathbf{X}^{(0)} \times \mathbf{Z} = \{(\mathbf{x}, \mathbf{z}_j) : s_i \in U \text{ and } S_j \in \mathbf{S}\}$, and define c_R and c_M as follows,*

$$c_R(\mathbf{x}, \mathbf{z}_j) = \begin{cases} 1 & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \qquad c_M(\mathbf{x}, \mathbf{z}_j) = \begin{cases} 1 & \text{if } s_i \notin S_j \\ 1/2 & \text{if } s_i \in S_j \end{cases}$$

Under this construction of the cost functions, each agent $\mathbf{x}$ will perform recourse if and only if $\mathbf{z}_i$ is revealed, and the disclosure probability $p = 1$. In the case that $\mathbf{z}_i$ is not revealed, the agent will elect to perform manipulation when any $\mathbf{z}_j$ is revealed where $j \neq i$ and $s_i \in S_j$. If neither criterion is met, the agent will elect to do nothing (remaining negatively classified). Combining these cases, we see that revealing each $\mathbf{z}_j$ causes exactly one agent to perform recourse, namely $\mathbf{x}_j$, and causes all $\mathbf{x}$ (with $s_i \in S_j$) to manipulate. Let $I = \{j_1, \ldots, j_k\}$ be the index set of the revealed features, then the number of agents manipulating is equal to $\left| \cup_{j \in I} S_j \right| - k$. Therefore providing k recourse actions to agents while minimizing the number of agents manipulating is equivalent to minimizing $\left| \cup_{j \in I} S_j \right|$.

C.3   Submodularity of the System's Utility

Theorem 7 *The system's objective function is* submodular *with respect to the size of the set of revealed features.*

Proof 2 *Given a revealed set $\mathbf{Z} \subseteq \mathbf{X}_R$, for agent with feature $\mathbf{x} \in \mathbf{X}_-$, let $S_m(\mathbf{x}, \mathbf{Z}) := \{z \in \mathbf{Z} : c_M(\mathbf{x}, z) \leq c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}, \mathbf{Z}))\}$ be the set of manipulation features that are cheaper than the minimum recourse action $\mathbf{z}_R(x, \mathbf{Z})$ given the revealed set $\mathbf{Z}$. Then the agent will perform recourse if and only if $S_m(\mathbf{x}, \mathbf{Z}) = \emptyset$. Given the cost function c_M and c_R, the principal can pre-compute each agent's manipulation set $S_m(\mathbf{x}, \mathbf{Z})$.*

The probability for the manipulation set $S_m(\mathbf{x}, \mathbf{Z})$ to overlap with a given revealed set $\mathbf{Z}$ is $P(\mathbf{x}; \mathbf{Z}) = \Pi_{z \in \mathbf{Z}_m(\mathbf{x}, \mathbf{Z})}(1 - p)$, where p is the disclosure probability for any criteria $\mathbf{z}$.

The goal for the system is to select a disclosure set $\mathbf{Z} \subseteq \mathbf{X}_R$ to minimize the overlap between $\mathbf{Z}$ and $S_m(\mathbf{x}, \mathbf{Z})$ for all agents, namely:

$$\min_{\mathbf{Z} \subseteq \mathbf{X}_R} u(\mathbf{Z}, \mathbf{X}_-) := \sum_{\mathbf{x} \in \mathbf{X}_-} \left(1 - P(\mathbf{x}; \mathbf{Z})\right) = \sum_{\mathbf{x} \in \mathbf{X}_-} \left(1 - \Pi_{z \in \mathbf{Z}_m(\mathbf{x}, \mathbf{Z})}(1 - p)\right) \tag{6}$$

To ease the notation, we use $u(\mathbf{Z})$ to shorthand $u(\mathbf{Z}, \mathbf{X}_-)$ since $\mathbf{X}_-$ is fixed in our setting. To show that Equation (6) is submodular, it is equivalent to prove that the objective function $u(\mathbf{Z}, \mathcal{X}_-)$ satisfies the diminishing returns property, which means $\forall A, B \subseteq \mathbf{Z}$ with $A \subseteq B \subseteq \mathbf{Z}$, and any criteria $z \in \mathbf{Z} \backslash B$, we want to show

$$u(A \cup \{z\}) - u(A) \geq u(B \cup \{z\}) - u(B)$$

Only four types of agents could potentially contribute to the marginal gain for U when the revealed sets are $A \cup \{z\}$ v.s. $B \cup \{z\}$:

1. when $S_m(\mathbf{x}, B \cup \{z\}) = B \cup \{z\}$

2. when $S_m(\mathbf{x}, B \cup \{z\}) = A \cup \{z\}$

3. when $S_m(\mathbf{x}, B \cup \{z\}) = \{z\}$

4. when $S_m(\mathbf{x}, B \cup \{z\}) = B \backslash A \cup \{z\}$

For the first three cases, we can verify that the two marginal gains are the same. For the last case, the two marginal gains are:

$$u(A + \{z\}) - u(A) = [1 - (1 - p)] - 0 = p$$
$$u(B + \{z\}) - u(B) = [1 - \Pi_{t \in \{B \backslash A \cup \{z\}\}}(1 - p)] - [1 - \Pi_{t \in \{B \backslash A\}}(1 - p)]$$
$$= p \times \Pi_{t \in \{B \backslash A\}}(1 - p)$$
$$\leq p$$

Since this holds for all agents, we show that adding a criterion z to a larger set B provides an equal or smaller marginal gain in the objective function compared to adding it to a smaller set A, satisfying the diminishing returns property. Therefore, the objective function defined in Equation (6) is submodular.

C.4 Proof for Theorem 1

Proof 3 *Notice that only agents $\mathbf{x} \in X^{(0)}$ who are originally negatively classified would request a recourse from the system in the first place, and both the recourse action and the manipulation actions that they are potentially going to take will be positively classified by the system. From the system's perspective, when the classifier is non-trivial (better than random guessing), all positively classified $\mathbf{x}$ are more likely to have true label $y = 1$, and all negatively classified $\mathbf{x}$ are more likely to have true label 0. When an agent with feature $\mathbf{x}$ takes recourse, the expected system utility change is:*

$$\Delta(\text{System's Utility})(\mathbf{x} \rightarrow \mathbf{z}_R) = \left(\mathbb{1}[y(\mathbf{z}_R) = 1, f(\mathbf{z}_R) = 1] - \mathbb{1}[y(\mathbf{z}_R) = -1, f(\mathbf{z}_R) = 1]\right) - 0$$
$$= 2\Pr[y(\mathbf{z}_R) = 1 | X = \mathbf{z}_R] - 1 \geq 0 \quad (\text{f is a non-trivial classifier, and } f(\mathbf{z}_R) = 1)$$

Similarly, when the agent takes manipulation, the expected system utility change is:

$$\Delta(\text{System's Utility})(\mathbf{x} \rightarrow \mathbf{z}_M) = \left(\mathbb{1}[y(\mathbf{x}) = 1, f(\mathbf{z}_M) = 1] - \mathbb{1}[y(\mathbf{x}) = -1, f(\mathbf{z}_M) = 1]\right) - 0$$
$$= 2\Pr[y(\mathbf{x}) = 1 | X = \mathbf{x}] - 1 \leq 0 \quad (\text{Since f is a non-trivial classifier, and } f(\mathbf{x}) = 0)$$

When the agent performs do-nothing, the system utility remains the same.

D   Proof for Section 6

We first prove a theorem on the monotonicity of social cost.

Theorem 2 (Monotonicity of Social Cost) *When the recourse cost $c_R(x, x')$ is monotonic in $\|x - x'\|$, and consider a linear threshold classifier. The social cost monotonically decreases in the easiest obtained recourse action.*

Proof 4 *Consider a 1-dimensional setting, where the system uses a linear threshold classifier $f(x) = \mathbb{1}[x \geq \tau]$. In this case, the optimal recourse action for any agent is always the minimum recourse action that has been revealed so far, namely $z_{\min} = \min_{z \in \boldsymbol{Z}} z$. Recall the definition of the social cost:*

$$cost(\boldsymbol{Z}, \boldsymbol{X}_-) = \sum_{\mathbf{x} \in \boldsymbol{X}_-} \left(c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}, \boldsymbol{Z}) - c_R(\mathbf{x}, \mathbf{x}_R) \right), where \; \mathbf{z}_R(\mathbf{x}, \boldsymbol{Z}) = \arg \min_{\mathbf{z} \in \boldsymbol{Z}} c_R(\mathbf{x}, \mathbf{z})$$

When the cost function is monotonic in the ℓ_2 norm, e.g., $c_R(x, x') = w_R \cdot \|x - x'\|$, we have

$$c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}, \boldsymbol{Z})) = w_R \cdot \|x - \mathbf{z}_R(\mathbf{x}, \boldsymbol{Z})\| = w_R \cdot \min_{z \in \boldsymbol{Z}} \|x - z\| = w_R \cdot \left(\min_{z \in \boldsymbol{Z}} z - x \right)$$

$$c_R(\mathbf{x}, \mathbf{x}_R) = w_R \cdot \|\mathbf{x} - \mathbf{x}_R\| = w_R \cdot \|\mathbf{x} - \tau\| = w_R \cdot (\tau - x)$$

Thus,

$$
\begin{aligned}
cost(\boldsymbol{Z}, \boldsymbol{X}_-) &= \sum_{\mathbf{x} \in \boldsymbol{X}_-} \left(c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}, \boldsymbol{Z})) - c_R(\mathbf{x}, \mathbf{z}_R) \right) \\
&= \sum_{\mathbf{x} \in \boldsymbol{X}_-} \left[w_R \cdot \left(\min_{z \in \boldsymbol{Z}} z - x \right) - w_R \cdot (\tau - x) \right] \\
&= |\boldsymbol{X}_-| \cdot w_R \cdot \left(\min_{z \in \boldsymbol{X}_-} z - \tau \right)
\end{aligned}
$$

As the size of $\boldsymbol{Z}$ gets larger (more recourse actions get revealed), $\min_{z \in \boldsymbol{Z}} z$ will be non-increasing, which means that $cost(\boldsymbol{Z}, \boldsymbol{X}_-)$ is monotonically decreasing.

E   Proofs for Section 7

E.1   Proof for Theorem 1

Proof 5 *Recall that given a revealed set $\boldsymbol{Z}$, with subsidy α, the corresponding recourse rate becomes:*

$$rec(\boldsymbol{Z}, \boldsymbol{X}_-; \alpha) = \frac{\sum_{\mathbf{x} \in \boldsymbol{X}_-} \mathbb{1} \left[\min_{\mathbf{z}' \in \boldsymbol{Z}} c_R(\mathbf{x}, \mathbf{z}'; \alpha) < \min \left(1, \; \min_{\mathbf{z}'' \in \boldsymbol{Z}} c_M(\mathbf{x}, \mathbf{z}'') \right) \right]}{|\boldsymbol{X}_-|}$$

In particular, with subsidy α, the cost of recourse becomes $(1 - \alpha) \cdot c_R(\mathbf{x}, \mathbf{z}')$, the cost of manipulation remains the same. Both optimal actions $\mathbf{z}_R$ and $\mathbf{z}_M$ remain the same.

Thus, for the nominator, we have:

$$\sum_{\mathbf{x} \in \mathbf{X}_-} \mathbb{1}\left[\min_{\mathbf{z}' \in \mathbf{Z}} c_R(\mathbf{x}, \mathbf{z}'; \alpha) \leq \min\left(1, \min_{\mathbf{z}'' \in \mathbf{Z}} c_M(\mathbf{x}, \mathbf{z}'')\right)\right]$$

$$= \sum_{\mathbf{x} \in \mathbf{X}_-} \mathbb{1}\left[\min_{\mathbf{z}' \in \mathbf{Z}} (1 - \alpha) \cdot c_R(\mathbf{x}, \mathbf{z}') \leq \min\left(1, \min_{\mathbf{z}'' \in \mathbf{Z}} c_M(\mathbf{x}, \mathbf{z}'')\right)\right]$$

$$= \sum_{\mathbf{x} \in \mathbf{X}_-} \mathbb{1}\left[(1 - \alpha) \cdot \underbrace{\min_{\mathbf{z}' \in \mathbf{Z}} c_R(\mathbf{x}, \mathbf{z}')}_{\text{fixed}} \leq \underbrace{\min\left(1, \min_{\mathbf{z}'' \in \mathbf{Z}} c_M(\mathbf{x}, \mathbf{z}'')\right)}_{\text{fixed}}\right]$$

$$= \sum_{\mathbf{x} \in \mathbf{X}_-} \mathbb{1}\left[(1 - \alpha) \cdot \underbrace{\min_{\mathbf{z}' \in \mathbf{Z}} c_R(\mathbf{x}, \mathbf{z}')}_{\text{fixed for a particular } x} \leq \underbrace{\min\left(1, \min_{\mathbf{z}'' \in \mathbf{Z}} c_M(\mathbf{x}, \mathbf{z}'')\right)}_{\text{fixed for a particular } x}\right]$$

$$= \sum_{\mathbf{x} \in \mathbf{X}_-} \mathbb{1}\left[\alpha \geq 1 - \underbrace{\frac{\min\left(1, \min_{\mathbf{z}'' \in \mathbf{Z}} c_M(\mathbf{x}, \mathbf{z}'')\right)}{\min_{\mathbf{z}' \in \mathbf{Z}} c_R(\mathbf{x}, \mathbf{z}')}}_{\text{fixed for a particular } x}\right]$$

As α becomes larger, this quantity will be non-decreasing. This implies that the recourse rate is a monotonically non-decreasing function of subsidy for a given revealed set $\mathbf{Z}$.

E.2   Proof for Theorem 2

Proof 6 *Again, consider a 1-dimensional setting, where the system uses a linear threshold classifier $f(x) = \mathbb{1}[x \geq \tau]$. In this case, the optimal recourse action for any agent is always the minimum recourse actions that has been revealed so far, namely $z_{\min} = \min_{z \in \mathbf{Z}} z$. Recall the definition of the social cost with subsidy level α:*

$$cost(\mathbf{Z}, \mathbf{X}_-; \alpha) = \sum_{\mathbf{x} \in \mathbf{X}_-} \left(c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}, \mathbf{Z}; \alpha); \alpha) - c_R(\mathbf{x}, \mathbf{z}_R)\right), \text{ where } \mathbf{z}_R(\mathbf{x}, \mathbf{Z}; \alpha) = \arg\min_{\mathbf{z} \in \mathbf{Z}} (1 - \alpha) c_R(\mathbf{x}, \mathbf{z})$$

In the 1-dimension case, we have

$$c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}, \mathbf{Z}; \alpha); \alpha) = (1 - \alpha) \cdot w_R \cdot \|\mathbf{x} - \mathbf{z}_R(\mathbf{x}, \mathbf{Z}; \alpha)\| = (1 - \alpha) \cdot w_R \cdot \min_{z \in \mathbf{Z}} \|x - z\| = (1 - \alpha) \cdot w_R \cdot \left(\min_{z \in \mathbf{Z}} z - x\right)$$

$$c_R(\mathbf{x}, \mathbf{z}_R; \alpha) = (1 - \alpha) \cdot w_R \cdot \|\mathbf{x} - \mathbf{z}_R\| = (1 - \alpha) \cdot w_R \cdot \|\mathbf{x} - \tau\| = (1 - \alpha) \cdot w_R \cdot (\tau - x)$$

Thus,

$$cost(\mathbf{Z}, \mathbf{X}_-; \alpha) = \sum_{\mathbf{x} \in \mathbf{X}_-} \left(c_R(\mathbf{x}, \mathbf{z}_R(\mathbf{x}; \mathbf{Z}; \alpha)) - c_R(\mathbf{x}, \mathbf{x}_R; \alpha)\right)$$

$$= \sum_{\mathbf{x} \in \mathbf{X}_-} \left[(1 - \alpha) \cdot w_R \cdot \left(\min_{z \in \mathbf{Z}} z - x\right) - (1 - \alpha) \cdot w_R \cdot (\tau - x)\right]$$

$$= (1 - \alpha) \cdot |\mathbf{X}_-| \cdot w_R \cdot \left(\min_{z \in \mathbf{Z}} z - \tau\right)$$

As the level of subsidy gets larger (α gets bigger, cheaper to perform recourse), $cost(\mathbf{Z}, \mathbf{X}_-; \alpha)$ will get smaller, which corresponds to a smaller social cost.

E.3   Proof for Theorem 3

Proof 1 *The system utility is defined as the difference between true positive and false positive after agent's actions. Let $\Pr[Y = 1|X = x]$ be the true qualification rate given a feature $X = x$, and assume it's also monotonic in X. u_0 is the system's initial utility (before providing recourse).*

Let the recourse region R_R and manipulation region R_M are defined as:

$$R_M = \{x \in \boldsymbol{X}^{(0)} : c_M(x, z_{\min}) < \min(1, c_R(x, z_{\min}))\}$$
$$R_R = \{x \in \boldsymbol{X}^{(0)} : c_R(x, z_{\min}) < \min(1, c_M(x, z_{\min}))\}$$

where $\mathcal{X}^{(0)}$ is the set of negatively classified agents. Then we have

System's utility$(z_{\min}) = \mathsf{TP} - \mathsf{FP}$

$$= u_0 + \underbrace{\int_{x \in R_M} \Pr(y = 1 | X = x) dx}_{\textit{TP from agents taking manipulation}} + \underbrace{\int_{x \in R_R} \Pr(y = 1 | X = z_{min}) dx}_{\textit{TP from agents taking recourse}}$$

$$- \underbrace{\int_{x \in R_M} (1 - \Pr(y = 1 | X = x)) \, dx}_{\textit{FP from agents taking manipulation}} - \underbrace{\int_{x \in R_R} (1 - \Pr(y = 1 | X = z_{min})) dx}_{\textit{FP from agents taking recourse}}$$

$$= u_0 + \int_{x \in R_M} (2 \cdot \Pr(y = 1 | X = x) - 1) \, dx + \int_{x \in R_R} (2 \Pr(y = 1 | X = z_{\min}) - 1) \, dx$$

$$= u_0 + \int_{x \in R_M} (2 \cdot \Pr(y = 1 | X = x) - 1) \, dx + (2 \Pr(y = 1 | X = z_{\min}) - 1) \int_{x \in R_R} dx$$

where $z_{\min} = \arg\min_{z \in \boldsymbol{Z}} z$ is the cheapest recourse actions.

Useful facts:

1. Suppose the classifier is a threshold classifier: $f = \mathbb{I}[x \geq \theta]$, we can further characterize the $\mathcal{X}^{(0)} = \{x \in \mathcal{X} : x \leq \theta\}$.

2. the minimum value of $z_{\min}$ is θ (the decision boundary).

3. Since $\Pr[y = 1 | X = x]$ is monotonic in x, $\forall x \in R_M, \Pr[y = 1 | X = x] \leq \Pr[y = 1 | X = \mathbf{z}_{min}]$

When we change the subsidy level α, the two regions change as:

$$\mathcal{X}_M^{(\alpha)} = \{x \in \boldsymbol{X}^{(0)} : c_M(x, z_{\min}) < \min(1, c_R(x, z_{\min}; \alpha))\}$$
$$R_R^{(\alpha)} = \{x \in \boldsymbol{X}^{(0)} : c_R(x, z_{\min}; \alpha) < \min(1, c_M(x, z_{\min}; \alpha))\}$$

where $c_R(x, x'; \alpha) = (1 - \alpha) \cdot c_R(x, x')$. As α becomes larger, we should expect $|\mathcal{X}_R^{(\alpha)}|$ to be larger and $|\mathcal{X}_M^{(\alpha)}|$ to be smaller.

When $c_R(x, x')$ and $c_M(x, x')$ are both monotonic in $\|x - x'\|$ and only cross once. wlog, assume

$$c_M(x, x') = \|x - x'\|, c_R(x, x'; \alpha) = \alpha \cdot w_R \cdot \|x - x'\| + b \quad (0 < w_R \leq 1, b < 1 \text{ to guarantee they only cross once})$$

we can further characterize the two regions:

$$\mathcal{X}_M^{(a)} = \{x : x \in [z_{\min} - \sqrt{\frac{b}{1 - \alpha \cdot w_R}}, \theta]\}, \mathcal{X}_R^{(a)} = \{x : x \in [z_{\min} - \sqrt{\frac{1 - b}{\alpha \cdot w_R}}, z_{\min} - \sqrt{\frac{b}{1 - \alpha \cdot w_R}}]\}$$

which gives us the size for the two regions as:

$$\left| \mathcal{X}_M^{(a)} \right| = \theta - z_{\min} + \sqrt{\frac{b}{1 - \alpha \cdot w_R}}, \quad \left| \mathcal{X}_R^{(a)} \right| = \sqrt{\frac{1 - b}{\alpha \cdot w_R}} - \sqrt{\frac{b}{1 - \alpha \cdot w_R}}$$

For $\alpha \in [0, 1]$, the rate in which the size of $\mathcal{X}_M^{(a)}$ and $\mathcal{X}_R^{(a)}$ changes as a function of the subsidy level α can be expressed as:

$$\frac{\partial |\mathcal{X}_M^{(a)}|}{\partial \alpha} = \frac{1}{2} \cdot b^{1/2} \cdot w \cdot (1 - aw)^{-3/2},$$

$$\frac{\partial |\mathcal{X}_R^{(a)}|}{\partial \alpha} = -\frac{1}{2} \sqrt{\frac{1 - b}{w}} \cdot a^{-3/2} - \frac{1}{2} \cdot b^{1/2} \cdot w \cdot (1 - aw)^{-3/2}$$

we can see the increase rate in the size of $R_R^{(\alpha)}$ is higher than the decrease rate in the size of $R_M^{(\alpha)}$. This, together with the fact that useful fact (3), tell us that the system's utility will be a monotonically increasing function in subsidy level α.

E.4 Proof for Theorem 5

Proof 7 *Recall from the proof for the recourse rate with subsidy, for a particular reveal set $\mathbf{Z}$ and a given set of negatively classified feature set $\mathbf{X}_-$, we have:*

$$rec(\mathbf{Z}, \mathbf{X}_-; \alpha) = \frac{\sum_{\mathbf{x} \in \mathbf{X}_-} \mathbf{1}\left[\alpha \geq 1 - \frac{\min\left(1, \min_{\mathbf{z}'' \in Z} c_M(\mathbf{x}, \mathbf{z}'')\right)}{\min_{\mathbf{z}' \in Z} c_R(\mathbf{x}, \mathbf{z}')}\right]}{|\mathbf{X}_-|}$$

To ease the notation, let's define $\gamma(x) = \frac{\sum_{\mathbf{x} \in \mathbf{X}_-} \mathbf{1}\left[\alpha \geq 1 - \frac{\min\left(1, \min_{\mathbf{z}'' \in Z} c_M(\mathbf{x}, \mathbf{z}'')\right)}{\min_{\mathbf{z}' \in Z} c_R(\mathbf{x}, \mathbf{z}')}\right]}{|\mathbf{X}_-|}$. Plug the expression into the definition for the disparity in recourse ratio for two groups g_0, g_1, we have:

$$Diff^{(rec)}(\mathbf{Z}, \mathbf{X}_-^{(g_0)}, \mathbf{X}_-^{(g_1)}) = \left| rec(\mathbf{Z}, \mathbf{X}_-^{(g_1)}, \alpha) - rec(\mathbf{Z}, \mathbf{X}_-^{(g_0)}, \alpha) \right|$$

$$= \left| \frac{\sum_{\mathbf{x} \in \mathbf{X}_-^{(g_1)}} \mathbb{1}\left[\alpha \geq 1 - \gamma(x)\right]}{|\mathbf{X}_-^{(g_1)}|} - \frac{\sum_{\mathbf{x} \in \mathbf{X}_-^{(g_0)}} \mathbb{1}\left[\alpha \geq 1 - \gamma(x)\right]}{|\mathbf{X}_-^{(g_0)}|} \right|$$

when the size of the two groups are similar, namely when $|\mathbf{X}_-^{(g_0)}| \approx |\mathbf{X}_-^{(g_1)}|$, we can roughly approximate the recourse difference by:

$$Diff^{(rec)}(\mathbf{Z}, \mathbf{X}_-^{(g_0)}, \mathbf{X}_-^{(g_1)}, \alpha) \approx \left| \sum_{\mathbf{x} \in \mathbf{X}_-^{(g_1)}} \mathbb{1}\left[\alpha \geq 1 - \gamma(x)\right] - \sum_{\mathbf{x} \in \mathbf{X}_-^{(g_0)}} \mathbb{1}\left[\alpha \geq 1 - \gamma(x)\right] \right|$$

We make the following observation:

- *When $\alpha = 0$: it corresponds to the situation where no subsidy is provided. This is the original disparity $Diff^{(rec)}(\mathbf{Z}, \mathbf{X}_-^{(g_0)}, \mathbf{X}_-^{(g_1)})$.*

- *When $\alpha = \alpha_{\max} = 1$, it corresponds to when the cost of recourse is 0, in this case, everyone takes recourse, which means the recourse difference is zero. Since $1 - \gamma(x) \leq 1 = \alpha_{\max}$ is also an upper bound on the value $1 = \gamma(x)$ for all $x \in \mathbf{X}_-$.*

For each group g_0 and g_1, if we rank x by their $1 - \gamma(x)$ value, then as we move α from 0 to 1, all the points that are to the left of the α will be counted towards $\mathbb{1}[\alpha \geq 1 - \gamma(x)]$. Thus the disparity will depend on the distribution of $1 - \gamma(x)$, which will mainly depend on the distribution of x, as well as the cost functions c_R and c_M. However, we are guaranteed to at least find an $1 < \alpha^* < 1$, such that after $\alpha > \alpha^*$, there is only one $x \in \mathbf{X}_-^{(g_0)}$ such that $\alpha \geq 1 - \gamma(x)$ is true. In this case, increasing α will only leads to decreasing in the disparity.

E.5 Proof for Theorem 4

Recall the statement of Theorem 4:

Theorem 3 (Subsidy Influence on Social Cost Disparity) *With subsidy α, the disparity in social cost for two group g_0, g_1 becomes:*

$$Diff^{(cost)}(\mathbf{Z}, \mathbf{X}^{(g_0)}, \mathbf{X}^{(g_1)}; \alpha) := \left| cost(\mathbf{Z}, \mathbf{X}_-^{(g_1)}; \alpha) - cost(\mathbf{Z}, \mathbf{X}_-^{(g_0)}; \alpha) \right|$$

Given a revealed set $\boldsymbol{Z}$, the social cost difference monotonically decreases in subsidies.

Proof 8 *Recall the definition of social cost difference:*

$$Diff^{(cost)}(S, \boldsymbol{X}_-^{(g_0)}, \boldsymbol{X}_-^{(g_1)}) := \left| cost(S, \boldsymbol{X}_-^{(g_1)}) - cost(S, \boldsymbol{X}_-^{(g_0)}) \right|$$

Again, consider a 1-dimensional setting, where the system uses a linear threshold classifier $f(x) = \mathbb{1}[x \geq \tau]$. In this case, the optimal recourse action for any agent is always the minimum recourse actions that has been revealed so far, namely $z_{\min} = \min_{z \in \boldsymbol{Z}} z$. Recall from the proof for social cost with subsidy, we have for a particular set $\mathcal{X}$:

$$cost(\boldsymbol{Z}, \boldsymbol{X}, \alpha) = (1 - \alpha) \cdot |\boldsymbol{X}| \cdot w_R \cdot \left(\min_{z \in \boldsymbol{Z}} z - \tau \right)$$

Plug it back to the definition of social cost difference at a certain subsidy level, we have:

$$
\begin{aligned}
Diff^{(cost)}(\boldsymbol{Z}, \boldsymbol{X}^{(g_0)}, \boldsymbol{X}^{(g_1)}; \alpha) &= \left| cost(\boldsymbol{Z}, \boldsymbol{X}^{(g_1)}) - cost(\boldsymbol{Z}, \boldsymbol{X}^{(g_0)}) \right| \\
&= \left| (1 - \alpha) \cdot |\boldsymbol{X}_-^{(g_0)}| \cdot (\min_{z \in \boldsymbol{Z}} -\tau) - (1 - \alpha) \cdot |\boldsymbol{X}_-^{(g_1)}| \cdot (\min_{z \in \boldsymbol{Z}} -\tau) \right| \\
&= \left| (1 - \alpha) \cdot (|\boldsymbol{X}_-^{(g_0)}| - |\boldsymbol{X}_-^{(g_1)}|) \cdot (\min_{z \in \boldsymbol{Z}} -\tau) \right|
\end{aligned}
$$

which is monotonically decreasing as α increases.

F   Additional Experimental Results

Additional Experimental Setup To optimize the system's utility and select the optimal set of features to reveal, we use the local search-based method provided in Orso et al. (2015).

Experiments Compute Resources All the experiments were run on a MacBook Pro with Apple M1 chip and 8GB memory. To finish 100 runs on the adult and law datasets for 5 different subsidies, it took roughly 3 hours; the German credit dataset will take roughly 5-6 hours.

F.1   Additional result on recourse rate difference between groups as a function of different values of subsidies

In Figure 4, we see the recourse rate difference between groups as a function of different values of subsidies. This figure serves to outline the parabolic nature relationship between subsidies and recourse rate difference. As mentioned previously, only those with already low recourse costs can benefit from subsidies for smaller subsides. Thus we see that smaller subsidies can initially result in greater disparity between agents, however, as subsides increases, they eventually decrease disparity to rates which are lower than the disparity without subsidies ($sub = 0$). Thus when deciding the amount of subsidies to choose, it is important for systems to be aware of the potential negative impacts (larger disparities between groups) that can result from smaller subsidies.

F.2   Additional Results Using Gradient Boosting Classifier

In this section, we present further empirical findings obtained by employing a Gradient Boosting Decision Tree as the training method. Overall, we observed similar behavior compared with training and logistic regression.
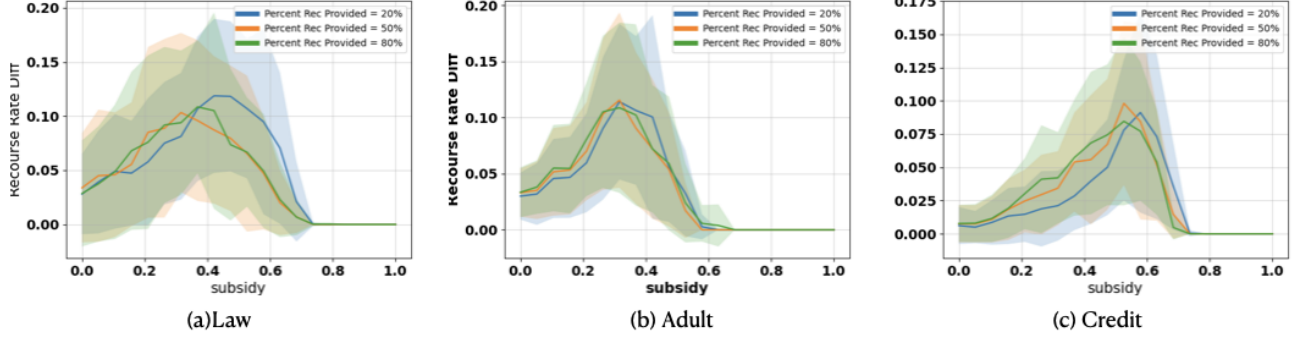
Figure 4: Recourse rate difference as a function of subsidy with 95% confidence intervals. Each line corresponds to a different percentage of the population with provided recourse actions.
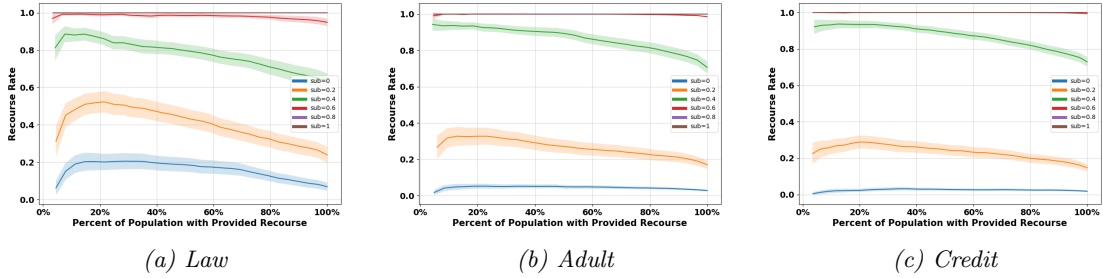

Figure 5: Fraction of the population performing recourse, with 95% confidence intervals. Each line corresponds to a different subsidy ratio "subs", i.e., the cost reduction applied to recourse.

G   Boarder Impact

Boarder Impact: By shedding light on the complex dynamics of recourse provision in automated systems, our paper challenges existing assumptions and reveals significant implications for both individuals and society as a whole. The identification of the natural tension between providing recourse and system exploitation highlights the delicate balance that must be maintained in algorithmic decision-making. This insight has profound consequences for fairness and equity, as strategic recourse withholding disproportionately affects vulnerable groups. Furthermore, the proposed framework offers a novel approach to analyzing the interplay of transparency, recourse, and manipulation, providing a valuable tool for future research in algorithmic fairness and accountability. The findings underscore the urgent need for policy interventions, such as recourse subsidies, to mitigate the adverse effects of system behavior on marginalized populations. Ultimately, this paper not only advances our theoretical understanding of algorithmic decision-making but also offers practical solutions to address the systemic biases inherent in automated systems.
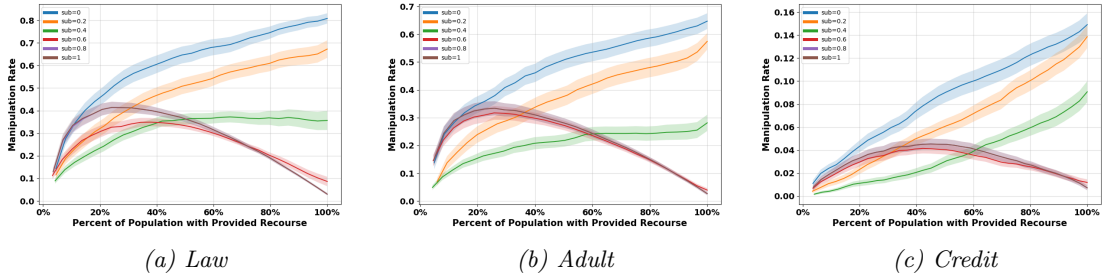

Figure 6: Fraction of the population performing manipulation, with 95% confidence intervals. Each line corresponds to a different subsidy ratio "subs", i.e., the cost reduction applied to recourse.
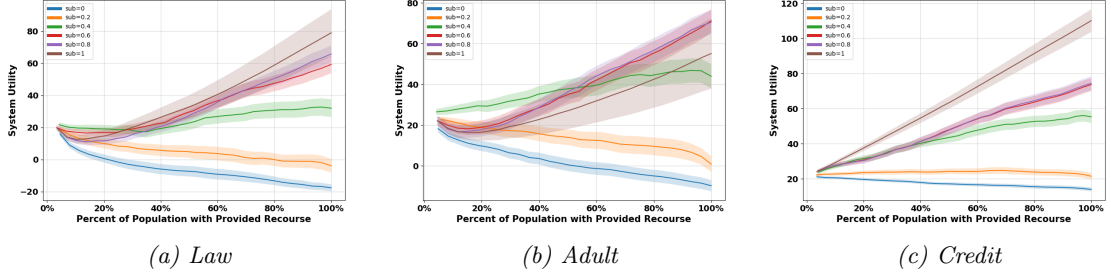
(a) Law

(b) Adult

(c) Credit

Figure 7: The system's utility as a function of the population percentage with provided recourse, with 95% confidence intervals. Each line corresponds to a different subsidy ratio "subs", i.e., the cost reduction applied to recourse.
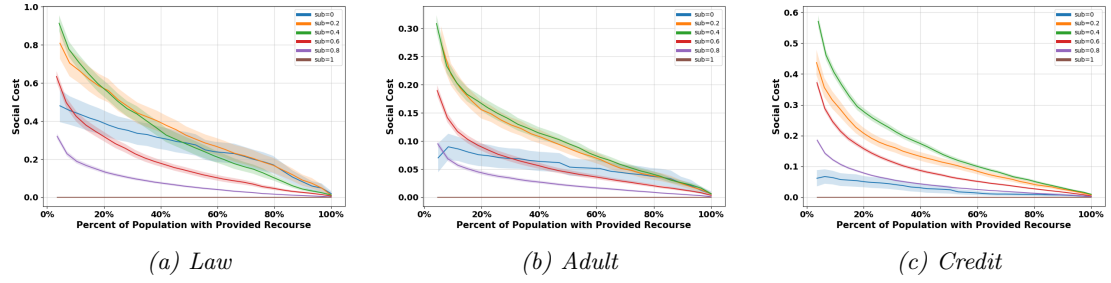

(a) Law

(b) Adult

(c) Credit

Figure 8: The social cost as a function of the population percentage with provided recourse, with 95% confidence intervals. Each line corresponds to a different subsidy ratio "subs", i.e., the cost reduction applied to recourse.

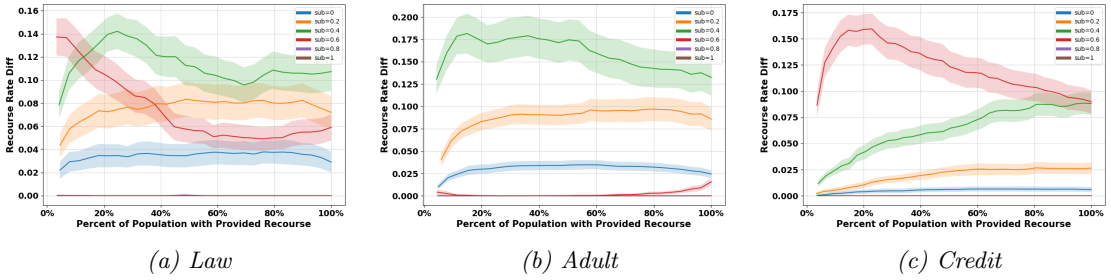

(a) Law

(b) Adult

(c) Credit

Figure 9: Difference in recourse rate between different sensitive attribute groups with 95% confidence intervals. Each line corresponds to a different subsidy ratio "subs", i.e., the cost reduction applied to recourse.
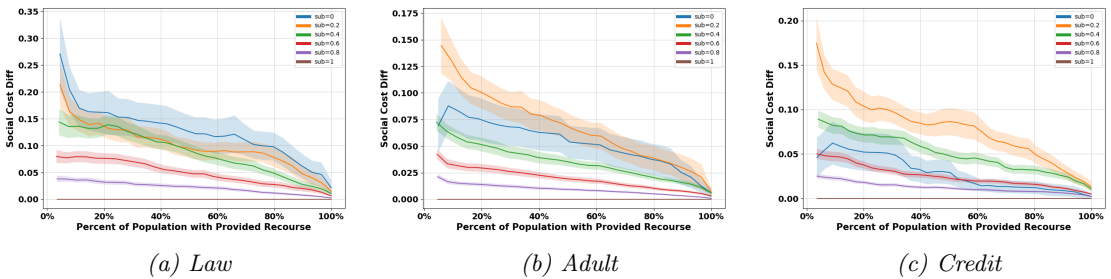

(a) Law

(b) Adult

(c) Credit

Figure 10: Difference in social cost between different sensitive attribute groups with 95% confidence intervals. Each line corresponds to a different subsidy ratio "subs", i.e., the cost reduction applied to recourse.

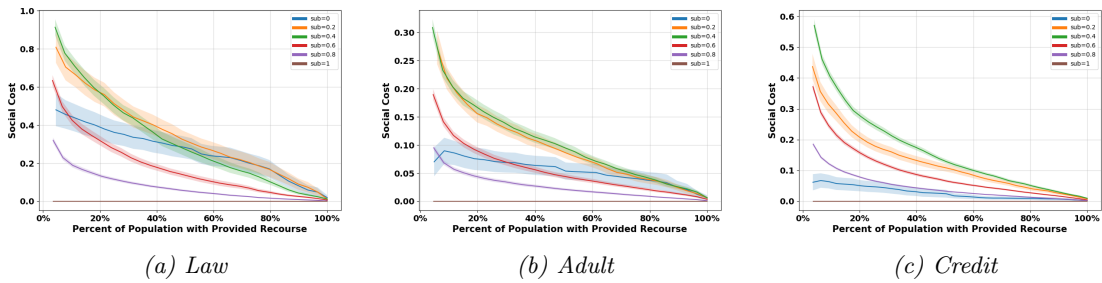(a) Law   (b) Adult   (c) Credit

Figure 11: The social cost as a function of the population percentage with provided recourse, with 95% confidence intervals. Each line corresponds to a different subsidy ratio "subs", i.e., the cost reduction applied to recourse.