Incentivizing Recourse through Auditing in Strategic Classification

Andrew Estornell[1] , **Yatong Chen**[2] , **Sanmay Das**[3] , **Yang Liu**[2] and **Yevgeniy Vorobeychik**[1]
[1]Washington University in Saint Louis
[2]University of California Santa Cruz
[3]George Mason University

Abstract

The increasing automation of high-stakes decisions with direct impact on the lives and well-being of individuals raises a number of important considerations. Prominent among these is strategic behavior by individuals hoping to achieve a more desirable outcome. Two forms of such behavior are commonly studied: 1) misreporting of individual attributes, and 2) recourse, or actions that truly change such attributes. The former involves deception, and is inherently undesirable, whereas the latter may well be a desirable goal insofar as it changes true individual qualification. We study misreporting and recourse as strategic choices by individuals within a unified framework. In particular, we propose auditing as a means to incentivize recourse actions over attribute manipulation, and characterize optimal audit policies for two types of principals, utility-maximizing and recourse-maximizing. Additionally, we consider subsidies as an incentive for recourse over manipulation, and show that even a utility-maximizing principal would be willing to devote a considerable amount of audit budget to providing such subsidies. Finally, we consider the problem of optimizing fines for failed audits, and bound the total cost incurred by the population as a result of audits.

1 Introduction

When the outcomes of algorithmic decision-making systems are consequential to those who interact with such systems, individuals receiving undesirable outcomes may take actions to improve their lot. For example, an individual whose credit card application has been denied may seek a means of improving their odds of approval on the next try. This issue has been studied from two perspectives: *strategic classification* and *actionable recourse*. Strategic classification involves *manipulation* of the reported features—that is, deception—typically resulting in an erroneous prediction [Hardt *et al.*, 2016], (e.g., an individual could inflate their income in order to qualify for a loan). As such, feature alterations in strategic classification are viewed as undesirable from a model-designer perspective; works in this area seek to dissuade individuals from performing manipulations, or reducing the impact that manipulations have on model performance. Actionable recourse, commonly termed *recourse*, on the other hand, entails agents *actually* changing their attributes and, consequently, the associated ground truth [Ustun *et al.*, 2019], (e.g., an individual getting a second job in order to meet the income requirements for a loan). In this case, the resulting change in the prediction is likely *correct*. Consequently, changes in features due to recourse tend to be viewed as beneficial even from the perspective of a principal interested in making decisions which maximize their own utility, maximize or social welfare. For example, by offering transparent recourse, a bank may be able to grant a larger number of profitable loans.

However, making recourse options transparent introduces a challenge for the principal, insofar as it is not necessarily evident whether attributes collected from individuals are a result of actual recourse actions, or manipulation that is facilitated by this added transparency. Consequently, there is a natural tension between offering transparent recourse options and at the same time being robust to strategic manipulations.

One way a principal can incentivize either truthful behavior or appropriate recourse is through the use of a (publicly announced) *audit policy*, where individuals are made aware that their declared attributes could be audited, and each individual may be subject to a penalty if found to be misrepresenting their true attributes [Blocki *et al.*, 2013]. While prior work has looked at the problem of optimal auditing to induce fully truthful behavior (i.e., incentive compatibility) [Estornell *et al.*, 2021], we initiate an investigation into the use of audits specifically in settings where agents have access to recourse.

Model: We model strategic classification in the presence of recourse as a game between a principal and a set of n agents where the principal has the ability to audit agents. As usual, each agent can be represented by an attribute (or feature) vector $\mathbf{x} \in \mathcal{X}$. There is a fixed, common-knowledge function $f : \mathcal{X} \to \mathbb{R}$ that represents the expected value to the principal of classifying as positive an agent represented by $\mathbf{x}$. The principal's goal is to positively classify only those with a positive expected value ($f(\mathbf{x}) \geq 0$) (a bank wishes to only issue loans to profitable customers). The central tension comes from the fact that agents can both (1) lie about their features, and (2) engage in costly actions that change their true features (the structure of costs is common knowledge). The principal's main tool to combat this is through audits. An audit con-

stitutes a check of whether an individual misrepresented their true feature vector. The principal can both impose fines and change the classification of agents found to be misreporting (e.g. individuals found to have over-reported their income are denied a credit card or loan, and may be subject to a fine).

Thus, we model the game as follows. The principal publicly declares an *audit policy*, which is a mapping from sets of n feature vectors to probabilities of auditing each of the n agents (n is common knowledge, and each agent's probability of being audited depends on the self-reported features of each agent in the whole set). A set of n agents each with private feature vectors $\mathbf{x}$ (unknown to the principal and other agents), arrives. Each agent then decides whether or not to perform recourse to change their feature vector, and what feature vector to report. Finally, the set of agents to be audited is decided based on the audit policy, and then the classifier induced by f is applied to the reports, less any agents caught misreporting, and a final decision is made on this population.

Contributions: In addition to introducing the modeling framework above that unifies manipulation and recourse, we obtain several consequential results in this model. We show that computing an optimal audit policy is tractable for both a utility-maximizing principal and a principal who simply wishes to maximize the number of agents choosing recourse (recourse-maximizing). This is true both when the costs of failing an audit are exogeneously specified and when the costs are chosen by the principal. We prove that when fines are exogeneously specified, the objectives of a recourse- and utility-maximizing principal are aligned for any distribution of agents, features, and cost of recourse.

We then turn our attention to studying a model of *subsidies*, where the principal can allot part of their audit budget to instead *subsidize agents to choose recourse*. We derive necessary and sufficient conditions for the principal to use a nonzero portion of their audit budget on subsidies. We show that even with subsidies, the objectives of a recourse- and utility-maximizing principal are again aligned when agents value positive classifications equally. We then characterize the relationship between auditing/subsidies, the total amount of fines or cost of recourse imposed on a population, and the fraction of individuals preferring recourse to manipulation.

2 Related Work

Recourse Recourse focuses on providing agents receiving undesirable outcomes from a model, with the ability to contest or improve their outcome via a modification to their attributes in a *genuine* manner (paying off debt to increase creditworthiness) [Ustun *et al.*, 2019; Karimi *et al.*, 2022; Karimi *et al.*, 2021; Upadhyay *et al.*, 2021; Gupta *et al.*, 2019; Venkatasubramanian and Alfano, 2020]. Our work makes use of the general formulation of recourse proposed in [Ustun *et al.*, 2019], which frames recourse as an optimization problem of finding *minimum* cost feature modifications which an agent can feasibly make and yield a desired outcome. Within this framework, we explore the role of auditing as a means of incentivizing recourse over manipulation.

Strategic Classification and Incentive Design Strategic Classification focuses on the problem of how to effectively make predictions in the presence of agents who behave strategically in order to obtain desirable outcomes. In particular, [Hardt *et al.*, 2016] first formalized strategic behavior in classification tasks as a sequential two-player game (i.e. a Stackelberg game) between a model designer and a strategic agent. In contrast to recourse, agent behavior is viewed as malicious in this context; model designers typically seek to disincentivize this behavior or limit its effects on model efficacy. Other similar formulations of this game have been studied [Levanon and Rosenfeld, 2021; Dong *et al.*, 2018; Milli *et al.*, 2019; Tsirtsis *et al.*, 2019]. Designing models which are robust to strategic behavior typically amounts to modifying the decision boundary of a classifier to be more selective [Hardt *et al.*, 2016; Milli *et al.*, 2019]. Shifts in the decision boundary of a classifier may increase agents' cost of recourse, leading to unnecessary cost for already qualified individuals. To avoid this issue in the context of recourse, we make use of auditing as a means of achieving model robustness (rather than modifying the classifier).

Another line of related work focuses on incentive design in the presence of strategic agents [Kleinberg and Raghavan, 2020; Chen *et al.*, 2020; Haghtalab *et al.*, 2020; Shavit *et al.*, 2020; Barsotti *et al.*, 2022], which aims to incentivize improving behaviors (e.g. recourse) and suppress pure manipulations. In particular, Barsotti *et al.* examines the relationship between model transparency and the level of manipulation detection which is required to promote improvements over manipulations. Rather than construct the detection mechanism, as is our focus, this work presumes the existence of a fixed mechanism which is independent of agents' choices.

Auditing Theory Auditing Theory examines problems in which a system (e.g., a bank) possesses the ability to verify (audit) information reported by an individual (e.g., a loan applicant). Auditing carries a negative consequence, such as a fine, when the reported information is found to be inauthentic. The work of [Blocki *et al.*, 2013] formulates auditing as a game between a defender attempting disincentivize manipulations, and an attacker attempting to avoid detection while obtaining a desired outcome (similar to a Stackelberg security game). Other works have studied auditing in the context of multiple individuals attempt to manipulate a classification or allocation system in order to gain a desired resource [Lundy *et al.*, 2019; Estornell *et al.*, 2021]. Auditing in the context of Strategic Classification remains relatively underexplored with the primary work being [Estornell *et al.*, 2021] which examines auditing as a means of inducing incentive compatibility (i.e. all agents truthfully report), but does not examine model robustness outside of this narrow lens. Works in this domain do not consider the ability for agents to perform recourse and are typically agnostic to system utility.

3 Model

We begin with a motivating example. A bank aims to maximize their expected profit by issuing fixed-rate credit cards (with set spending limits and interest rates). Because of the high volume compared with, say, corporate loans, credit cards are a major area where banks use algorithmic decision-making [Butaru *et al.*, 2016]. Each applicant (with applica-

tion $\mathbf{x}$) is approved for a fixed-rate card if the bank's model predicts that an applicant will offer a positive profit. While the profitability comes from different channels, e.g. building a relationship with a client who will then use the bank for other services versus actual interest payments, the main risk in issuing a card is that the customer will default after running up a balance [Khandani *et al.*, 2010], so banks want to filter out those applicants. IF the bank denies an application, the bank's utility is 0 as no money is exchanged. The bank may offer denied applicants access to recourse, i.e., a plan for making the applicant more creditworthy, such as paying off outstanding debt or increasing income. However, when applicants have knowledge of recourse actions, they may report that they have taken such actions in order to get approved, without actually taking the actions (e.g. hiding debt or inflating income). The bank could audit applicants by verifying information in their applications. However, since this is costly, the audit budget is limited.

We now present our formal model of auditing and recourse. Let $\mathcal{D}$ be a distribution over features $\mathcal{X} \subset \mathbb{R}^d$ with probability measure p. Consider a principal who aims to make a binary decision $\hat{y}(\mathbf{x}) \in \{0, 1\}$ for each input feature vector $\mathbf{x}$, for example, to approve or deny a loan. We refer to the decision $\hat{y}(\mathbf{x}) = 1$ as *selection*, with $\hat{y}(\mathbf{x}) = 0$ corresponding to $\mathbf{x}$ not being selected. For any *actual* feature vector $\mathbf{x}$ (to distinguish from manipulated features we discuss below), the principal receives a utility of $u_p(\mathbf{x})$ whenever $\hat{y}(\mathbf{x}) = 1$ (e.g., expected profit from a loan) and utility of 0 otherwise; in other words, the principal's utility is $u_p(\mathbf{x})\hat{y}(\mathbf{x})$.

Prediction function We assume that the principal's utility from selecting $\mathbf{x}$ is based on an objective measure, such as loan repayment rate, that is not known directly, but can be estimated from data. Thus, let $f : \mathcal{X} \to \mathbb{R}$ be a model learned from data that predicts $u_p(\mathbf{x})$. For example, f can predict the probability that a loan is repaid, multiplied by expected profits conditional on repayment. Importantly, we assume that f is fixed and common-knowledge, and is applied to the *reported* features. The application of f is thus mechanistic and not an action under the control of the principal in the game-theoretic sense. This is consistent with our use cases – bank regulators, for example typically require that a model is demonstrably a valid predictor and that it should be used consistently across the entire population of applicants for a period of time. Thus, f is simply used to select all $\mathbf{x}$ that yield a predicted utility above a given threshold θ:

$$\hat{y}(\mathbf{x}) = \mathbb{I}\big[f(\mathbf{x}) \geq \theta\big].$$

If we set $\theta = 0$, this has the natural interpretation in the context of loans that all applications with positive expected utility (based on the reported features) are approved.

Principal's Actions: Auditing Agents can misreport their feature vectors. The principal's main tool to disincentivize such misrepresentation is the use of audits. When the principal audits an agent reporting features $\mathbf{x}'$, the agent's true features $\mathbf{x}$ are revealed to the principal. Failing an audit, i.e., being audited when $\mathbf{x}' \neq \mathbf{x}$ will result in the agent paying a fine; we follow the models of auditing in [Blocki *et al.*, 2013; Estornell *et al.*, 2021] and assume agents pay a constant fine C when they are caught manipulating, in addition to not being selected. Before agents report their features, the principal publicly declares its *audit policy*.

Definition 1. (Audit Policy) *Given a set of n agents with true features $\mathbf{X}$ and reported features $\mathbf{X}'$, an* audit policy *is a mapping $\alpha : \mathcal{X}^{n+1} \to [0, 1]$ where $\alpha(\mathbf{x}'; \mathbf{X}')$ corresponds to the probability that an agent reporting features $\mathbf{x}'$ is audited, given the set of reports $\mathbf{X}'$ for the n agents. The principal is limited B audits on average, i.e., $\mathbb{E}\big[\sum_{\mathbf{x}' \in \mathbf{X}'} \alpha(\mathbf{x}'; \mathbf{X}')\big] \leq B$.*

An audit of a particular agent is a check whether $\mathbb{I}[\mathbf{x}' \neq \mathbf{x}]$, which we assume to be reliable. Agents caught misreporting their features are subject to a fine $C \in \mathbb{R}_{\geq 0}$.

Agents An agent with true features $\mathbf{x}$ gains utility $u_a(\mathbf{x})$ when approved by the principal, and 0 otherwise. When reporting features $\mathbf{x}'$, and not being caught by an audit, the agent then obtains utility $u_a(\mathbf{x})\hat{y}(\mathbf{x}')$. In addition to the general case, we also consider a special case where the utility of being selected is a constant, i.e., $u_a(\mathbf{x}) = \bar{u}_a$ for all $\mathbf{x}$. This special case has received most attention in prior literature, particularly in the context of recourse [Ustun *et al.*, 2019].

Agents' Actions: Recourse and Manipulation Formally, n agents arrive i.i.d. with features $\mathbf{x} \sim \mathcal{D}$; we assume that $\mathcal{D}$ is common knowledge. We use $\mathbf{X} \sim \mathcal{D}$ to indicate a collection of n feature vectors thereby generated. Each agent has an action space comprised of two qualitatively distinct types of actions: recourse and manipulation. We allow arbitrary composition of these, although prove below that such compositions are dominated by a choice of manipulation, recourse, or neither (reporting true initial features $\mathbf{x}$). Let $\mathbf{z}$ denote a recourse choice, which we restrict to be in the set $A(\mathbf{x})$ that defines what is *actionable* [Ustun *et al.*, 2019]. The agent always has the option to do nothing, i.e. $\mathbf{x} \in A(\mathbf{x})$, and if the agent elects this do-nothing action (which carries no cost), then $\mathbf{z} = \mathbf{x}$. The cost of a recourse action $\mathbf{z}$ for an agent with initial features $\mathbf{x}$ is denoted by $c_R(\mathbf{x}, \mathbf{z})$. We use $\mathbf{z}'$ to denote reported (potentially manipulated) features.

While selection decisions $\hat{y}$ are implemented independently for each reported feature vector $\mathbf{z}'$, the audit policy $\alpha(\mathbf{z}'; \mathbf{Z}')$ depends on the full collection of n *reported* feature vectors of all agents, namely $\mathbf{Z}'$. Let $g(\mathbf{x})$ be the strategy of an agent with true features $\mathbf{x}$ in the choice of both recourse $\mathbf{z}$ and reported (and possibly untruthful) features $\mathbf{z}'$. We restrict attention to symmetric pure strategies, so that g deterministically returns a pair $(\mathbf{z}, \mathbf{z}')$. Given a symmetric strategy profile g and an agent who reports a feature vector $\mathbf{z}'$, the probability of this agent being audited is $\mathbb{E}\big[\alpha(\mathbf{z}'; g(\mathbf{X}))\big]$, where the expectation is with respect to $\mathbf{X} \sim \mathcal{D}$ (here, it is only the final reports induced by g that matter). We define the expected cost of manipulation for an agent with true features $\mathbf{z}$ (possibly after recourse) and reported features $\mathbf{z}'$, when all other agents jointly follow strategy g as

$$c_A(\mathbf{z}, \mathbf{z}'; g) = \mathbb{E}\big[\alpha(\mathbf{z}'; g(\mathbf{X}))\big]\mathbb{I}\big[\mathbf{z}' \neq \mathbf{z}\big]\big(u_a(\mathbf{z})\hat{y}(\mathbf{z}') + C\big).$$

Putting everything together, the expected utility of an agent with initial features $\mathbf{x}$, recourse $\mathbf{z}$, and reported features $\mathbf{z}'$, given a symmetric strategy profile g followed by all others, is

$$U_a(\mathbf{z}, \mathbf{z}', g; \mathbf{x}) = u_a(\mathbf{z})\hat{y}(\mathbf{z}') - c_R(\mathbf{x}, \mathbf{z}) - c_A(\mathbf{z}, \mathbf{z}'; g). \quad (1)$$

When all agents follow g, we simply write $U_a(g;\mathbf{x})$, as $(\mathbf{z},\mathbf{z}') = g(\mathbf{x})$. Our solution concept for agent strategies is a (pure-strategy symmetric) Bayes-Nash equilibrium.

Definition 2. *A symmetric pure-strategy strategy profile g is a* Bayes-Nash equilibrium *(BNE) if for all agents i with initial features $\mathbf{x}_i$, the action $g(\mathbf{x}_i)$ is a best response, i.e., $U_a(g;\mathbf{x}_i) \geq U_a(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}'_i, g;\mathbf{x}_i)$ for all $\bar{\mathbf{z}}_i \in A(\mathbf{x}_i)$ and $\bar{\mathbf{z}}'_i$. We denote the BNE profile with the maximum number of manipulations as $g_{\max}$.*

4 Optimal Auditing

In this section we investigate the audit polices of both a recourse-maximizing principal and a utility-maximizing principal. We begin by characterizing some key facts about agents' best responses given the principal's audit policy.

Lemma 1. *It is never a best response for an agent to perform both recourse and manipulation i.e. either $\mathbf{z} = \mathbf{x}$ or $\mathbf{z} = \mathbf{z}'$.*

This result follows from the fact that agent utility is independent of the report $\mathbf{z}'$ whenever $\hat{y}(\mathbf{z}') = 1$; the full proof is in Section A.1 of the Supplement.

Next we examine the best response of each agent $\mathbf{x}$, with recourse action $\mathbf{z}$ ($\mathbf{z} = \mathbf{x}$ if no recourse occurs), given prediction function f, decision making scheme $\hat{y}$, audit policy α, and fine C. For any strategy g by other agents, the optimal manipulation and recourse are given respectively by,

$$\mathbf{x}_M = \arg\max_{\mathbf{z}' \neq \mathbf{x}} \; u_a(\mathbf{x}) - c_A(\mathbf{x},\mathbf{z}';g) \quad \text{s.t. } \hat{y}(\mathbf{z}') = 1 \quad (2)$$

$$\mathbf{x}_R = \arg\max_{\mathbf{z} \in A(\mathbf{x})} \; u_a(\mathbf{z}) - c_R(\mathbf{x},\mathbf{z}) \quad \text{s.t. } \hat{y}(\mathbf{z}) = 1. \quad (3)$$

For an agent $\mathbf{x}$, let $U_{a,R}(\mathbf{x}) = u_a(\mathbf{x}_R) - c_R(\mathbf{x},\mathbf{x}_R)$ and $U_{a,M}(\mathbf{x}) = u_a(\mathbf{x}) - c_A(\mathbf{x},\mathbf{x}_M;g)$, i.e. the agent's respective utility gain from recourse or manipulation. The next lemma characterizes the structure of agent best response actions in terms of their expected utility gain.

Lemma 2. *The best response of an agent with features $\mathbf{x}$ has the following form:*

$$\mathbf{z}^* = \begin{cases} \mathbf{x} & \text{if } \hat{y}(\mathbf{x}) = 1 & (4a) \\ \mathbf{x}_R & \text{if } U_{a,R}(\mathbf{x}) \geq \max(0, U_{a,M}(\mathbf{x})) & (4b) \\ \mathbf{x}_M & \text{if } U_{a,M}(\mathbf{x}) \geq \max(0, U_{a,R}(\mathbf{x})) & (4c) \\ \mathbf{x} & \text{otherwise} & (4d) \end{cases}$$

where Equations 4a, 4d correspond to truthful reporting, Equation 4b corresponds to recourse, and Equation 4c corresponds to manipulation.

Lemma 2 follows directly from each action's definition.

Theorem 1. *Let $g_{\max}$ be the BNE profile which has the maximum number manipulations. If an audit policy α is recourse (or utility) maximizing with respect to $g_{\max}$, it is recourse (or utility) maximizing for any other BNE profile g.*

The proof of Theorem 1 is deferred to Section A.1 of the Supplement. The intuition is that the efficacy of any audit policy α is monotonically decreasing in the number of agents who manipulate. Henceforth, we leverage this result to only consider the principal's objective with respect to $g_{\max}$.

Next we formalize the objective of the principal. We consider two types of principals: a population-oriented principal who aims to maximize the proportion of agents that prefer recourse to manipulation (dubbed *recourse-maximizing*) and a principal who aims to maximize the total utility gain of the decisions made by $\hat{y}$ (dubbed *utility-maximizing*). These objectives respectively represent a principal who is socially-oriented (we treat recourse as a kind of social good, as it benefits participants), or solely self interested.

Definition 3. ***(Recourse-Maximizing Principal)**: A principal is* recourse-maximizing *if their objective is to select an audit policy α which maximizes the proportion of agents who prefer recourse over manipulation:*

$$\alpha^* = \arg\max_{\alpha} \mathbb{P}_{\mathbf{X}}\big(U_{a,R}(\mathbf{x}) \geq U_{a,M}(\mathbf{x})\big) \quad (5)$$
$$\text{s.t. } \mathbb{E}_\alpha\Big[\sum_{\mathbf{z}' \in \mathbf{Z}'} \alpha(\mathbf{z}';\mathbf{Z}') | \mathbf{Z}'\Big] \leq B \quad \forall\, \mathbf{Z}'$$

Definition 4. ***(Utility Maximizing Principal)**: A principal is* utility maximizing *if their objective is to select an audit policy α which maximizes the principal's utility. For an agent with true features $\mathbf{x}$, let $\mathbf{z} = \mathbf{x}_R$ if the agent performs recourse and $\mathbf{z} = \mathbf{x}$ otherwise, and let $\mathbf{z}'$ be the agent's report. This objective can be framed as,*

$$\alpha^* = \arg\max_{\alpha} \mathbb{E}\big[\hat{y}(\mathbf{z}') f(\mathbf{z}) \big(\alpha(\mathbf{z}';\mathbf{Z}')\mathbb{I}[\mathbf{z} \neq \mathbf{z}'] + \mathbb{I}[\mathbf{z} = \mathbf{z}']\big)\big] \quad (6)$$
$$\text{s.t. } \mathbb{E}_\alpha\Big[\sum_{\mathbf{z}' \in \mathbf{Z}'} \alpha(\mathbf{z}';\mathbf{Z}') | \mathbf{Z}'\Big] \leq B \quad \forall\, \mathbf{Z}'$$

We now turn to the characterization of optimal audit policies. We demonstrate the somewhat counter-intuitive result that for both a recourse-maximizing and utility-maximizing principal, uniformly auditing all positively classified agents is optimal under any distribution of agent types, recourse cost function, prediction model, and agent utility function.

Theorem 2. *For any recourse cost function $c_R(\mathbf{x},\mathbf{z})$, agent utility function $u_a(\mathbf{x})$, feature distribution $\mathcal{D}$, a recourse maximizing principal with budget B and fine C has optimal policy*

$$\alpha(\mathbf{z}';\mathbf{Z}') = B/|\mathbf{Z}'^{(1)}|, \quad \forall\, \mathbf{z}' \in \mathbf{Z}'^{(1)}, \forall\, \mathbf{X}$$

when agent reports $\mathbf{Z}'$ are induced the BNE profile $g_{\max}$; $\mathbf{Z}'^{(1)}$ is the set of all reports $\mathbf{z}'$ with $\hat{y}(\mathbf{z}') = 1$.

Proof Sketch. The full proof is deferred to Section A.1 of the Supplement. A recourse maximizing principal aims to select an audit policy α in order to maximize the expected number of agents performing recourse. Only agent with $\hat{y}(\mathbf{x}) = 0$ have incentive to perform recourse or manipulation, and will select the actions with highest expected payoff. Despite the fact that the selection of α depends on agents' actions, which themselves depend on α as well as the actions of other agents, we can express the condition that an agent with true feature $\mathbf{x}$ will not manipulate as,

$$\frac{u_a(\mathbf{x}) - u_a(\mathbf{x}_R) + c_R(\mathbf{x},\mathbf{x}_R)}{u_a(\mathbf{x}) + C} \leq \min_{\mathbf{z}' \in \mathcal{X}^{(1)}} \mathbb{E}_{\mathbf{X}}\big[\alpha(\mathbf{z}';g(\mathbf{X}))\big] \quad (7)$$

A recourse maximizing principal aims to select α such that the above condition holds for as large a fraction of agents as is

possible in expectation. Since the left-hand side of Inequality 7, is independent of α and g, and the right-hand side is a minimization over *all* features in $\mathcal{X}^{(1)}$, the optimal solution has $\alpha(\mathbf{z}'_1; g(\mathbf{X})) = \alpha(\mathbf{z}'_2; g(\mathbf{X}))$ for all $\mathbf{z}'_1, \mathbf{z}'_2 \in \mathbf{X}'^{(1)}$, i.e. $\alpha(\mathbf{z}'; g(\mathbf{X})) = B/|\mathbf{X}'^{(1)}|$ for all $\mathbf{z}' \in \mathbf{X}'^{(1)}$. □

Theorem 3. *For any recourse cost function $c_R(\mathbf{x}, \mathbf{z})$, agent utility function $u_a(\mathbf{x})$, feature distribution $\mathcal{D}$, the policy in Theorem 2 (uniform auditing) is a utility maximizing policy, when the induced BNE profile of agents is $g_{\max}$.*

Proof sketch. The full proof of this theorem is deferred to Section A.1 of the Supplement. The proof strategy is similar to that of Theorem 2; we can formulate the principal's objective as maximizing the minimum $\alpha(\mathbf{z}'; g(\mathbf{X}))$ such that a condition similar to Inequality 7 holds for the largest fraction of agents. The key difference being that not all agents contribute equality to the principal's objective; agent $\mathbf{x}$ yields utility $f(\mathbf{x})$. However, this formulation requires only two observations: 1.) an agent $\mathbf{x}$ with $\hat{y}(\mathbf{x}) = 0$ ($\mathbf{x}$ has incentive to manipulate) yields negative utility, and 2.) agents performing recourse offers nonnegative utility, which is at least as good as successfully auditing a manipulation (utility 0). Recourse (manipulations) offers nonnegative (nonpositive) utility, and the principal's utility is monotone increasing in the number of agents performing recourse. Combining these observations, the principal's objective becomes precisely recourse maximization, and the proof follows from Theorem 2. □

Remark: *Theorems 2 and 3 show an equivalence between a recourse-maximizing and utility-maximizing principal. The significance of which is threefold: (1) the actions of a self-interested (utility-maximizing) principal are as beneficial to the population as the actions of a recourse-maximizing principal directly trying to maximize for population benefit, (2) self-interested auditing decreases the percentage of agents which engage in "risky" and potentially socially detrimental behavior (manipulation), and (3) optimal auditing* does not *require any knowledge of dynamics of agents recourse actions (e.g. solving Program 3, or even knowing c_R).*

5 Auditing With Subsides

Audits provide a punitive measure for incentivizing recourse over manipulation. Another natural option is to offer subsidies that make recourse cheaper to implement for agents. Here we investigate how the principal optimally splits the limited budget between auditing and subsidies. For example, a bank may choose to allocate a fraction of their budget from application verification to the development of educational material to help increase financial literacy. Our key result is that in the important special case of constant utilities, both recourse-maximizing and (own) utility-maximizing principals choose the same fraction of budget for subsidies. Moreover, we show that despite the complex interdependencies of the problem, when agent utilities are constant, the objective of both principals can be formulated as a single-dimensional optimization problem, depending only on the impact of subsidies on the cost of recourse and audit budget. We begin by formalizing subsidies in our model.

Definition 5. *A* subsidy function $s : [0, B] \to [0, 1]$ *yields a multiplicative decrease in the cost of recourse, such that for a subsidy budget b, the cost of recourse becomes $s(b)c_R(\mathbf{x}, \mathbf{z})$, and the remaining budget $B - b$ is then used for auditing. Subsidy functions $s(b)$ are decreasing in b and $s(0) = 1$ (allocating no subsidies recovers the original recourse cost).*

Remark: *For any subsidy trade-off b^* with $s(b^*) = 0$, the cost of recourse is $s(b^*)c_R(\mathbf{x}, \mathbf{z}) = 0$ for all $\mathbf{x}, \mathbf{z}$. When such a trade-off exists, it is always optimal for the principal to select b^* as their subsidy allocation (i.e., their objective reduces to univariate root finding of $s(b)$). Consequently, we henceforth assume that $s(b) > 0$.*

Next, we present our key result showing that when agent utilities are constant, optimal subsidy characterization is identical for either recourse- or utility-maximizing principal, and amounts to solving a one-dimensional optimization problem.

Theorem 4. *Suppose that agent utilities are constant, i.e., $u_a(\mathbf{x}_i) = \bar{u}_a$, and the induced BNE profile of agents is $g_{\max}$. Then, for both a recourse-maximizing and utility-maximizing principal, the optimal subsidy is given by*

$$b^* = \arg \max_{b \in [0,B]} \frac{B - b}{s(b)} \tag{8}$$

Proof Sketch . The proof is deferred Section A.2 of the Supplement. The strategy for this proof is again to induce an ordering on agents via the difference in their expected utility gain from either recourse or manipulation. The key challenge in the case of subsidies is that the subsidy function $s(b)$ affects this difference in utility. In the case of constant agent utility the optimal recourse action $\mathbf{x}_R$, for agent $\mathbf{x}$, is invariant w.r.t. b (although the cost changes w.r.t. b). For any fixed b, uniformly auditing with the remaining $B - b$ budget is optimal. Thus, the condition that agent $\mathbf{x}$ prefers recourse is

$$\frac{c_R(\mathbf{x}, \mathbf{x}_R)}{\left(C + \bar{u}_a\right) \mathbb{E}_{\mathbf{X}}\left[1/|\mathbf{X}^{(1)}|\right]} \leq \frac{B - b}{s(b)},$$

Similar to Theorem 2, the left-hand side is independent of b, and maximizing $(B - b)/s(b)$ maximizes the number of agents preferring recourse. The argument for utility maximization follows a similar line of reason to Theorem 3. □

Illustration: *To gain some intuition into the result of Theorem 4, consider $s(b) = \frac{1}{b+1}$, where the impact of subsidies on recourse costs exhibits diminishing returns in the subsidy allocation. In this case, the objective can be solved analytically, obtaining the optimal subsidy $b^* = \frac{B-1}{2}$. Thus, the principal, whether maximizing overall welfare or their own utility, would allocate nearly half of the audit budget to subsidies. The reason is that even a self-interested principal actually benefits from providing subsidies and thereby incentivizing recourse, as such actions also increase the principal's profits, whereas manipulation results in an expected loss.*

Corollary 1. *When agent utility is constant, both a recourse-maximizing and utility-maximizing principal will allot a nonzero portion of their budget to subsides if and only if there exists some b s.t. $s(b) \leq 1 - b/B$, i.e. s has better than linear scaling for at least one value of b.*

In contrast to the case of constant agent utilities, however, optimal subsidy becomes non-trivial for general agent utilities. Moreover, the alignment between recourse- and utility-maximizing principal no longer obtains.

Theorem 5. *For general agent utilities, recourse maximization and utility maximization are no longer aligned.*

The full proof, and further details on subsides for general agent utility, are provided in the Supplement Section A.2.

6 Optimal Fines

Thus far, our analysis has assumed exogenously specified fines C from failing an audit. We now consider a principal who jointly optimizes α and C. We adopt the model of Blocki *et al.* [2013] in which the principal can select the value of C and suffers a cost $L(C)$ as a result, which can capture public concerns about unfairly high penalties, regulatory pressures, etc. For ease of exposition we restrict attention to $L(C) = \gamma C$ for exogenously specified γ; however, our results hold for an arbitrary monotone function $L(C)$. The optimal fine and audit policy can then be found via

$$\max_{\alpha, C} \; V(C, \alpha) \text{ - } \gamma C, \qquad \text{s.t. } \mathbb{E}_{\alpha}\big[\sum_{\mathbf{z}' \in \mathbf{Z}'} \alpha(\mathbf{z}'; \mathbf{Z}') | \; \mathbf{Z}' \big] \leq B$$

where $V(C, \alpha)$ is the principal's utility (Equations (5), (6)). In Section A.3 of the Supplement we discuss the alignment of these objectives. Here, we show that for constant agent utility, uniform auditing is again optimal and the optimal fine can be formulated as a one-dimensional optimization problem solely in terms of the CDF of $c_R(\mathbf{x}, \mathbf{x}_R)$ and the penalty on C.

Theorem 6. *Let F_R be the CDF of the cost of recourse, i.e., $F_R(k) = \mathbb{P}_{\mathbf{x}}\big(c_R(\mathbf{x}, \mathbf{x}_R) \leq k\big)$. Suppose agent utility is constant, $u_a(\mathbf{x}) = \bar{u}_a$, and the induced BNE profile of agents is $g_{\max}$. Then uniform auditing is optimal for both recourse and utility maximization and the optimal fine is,*

$$C^* = \arg\max_{C} \; nF_R\big(B(C + \bar{u}_a)/n\big) - \gamma C$$

The full proof is provided the Supplement, Section A.3. This result again demonstrates the ease of optimal auditing, as the principal's objective reduces entirely to finding C^*. Although the difference of two monotonic functions (the CDF and the penalty) may be intractable to optimize in general, there exists a wide array of technique for optimizing such functions efficiently in practice [Sergeyev, 1998; Locatelli, 1997; Hansen, 1979]. If F_R is Lipschitz smooth in C, then these methods yield arbitrarily good approximations of C^*, this holds for any Lipschitz smooth monotonic $L(C)$.

7 Costs of Auditing to the Population

In domains where recourse is a salient consideration, it is natural to examine the average cost suffered by a population when performing recourse [Ustun *et al.*, 2019]. With the introduction of auditing and subsides into such domains, it becomes imperative to consider costs/fines imposed on the population as both a function of auditing and subsides.

We first describe the differences between the impact on the utility of the principal and that of the agents. In particular, as the auditing budget B and fine C increase, the principal's utility gain is monotonically increasing, while the agents' utility gain is monotonically decreasing.

Theorem 7. *Average agent utility is monotonically decreasing in B and C. In contrast, the principal's expected utility is monotonically increasing in B and C.*

Proof Sketch. The expected utility of and agent $\mathbf{x}$ misreporting $\mathbf{z}'$ can be expressed as

$$u_a(\mathbf{x}) - \mathbb{E}\big[B/|\mathbf{Z}'^{(1)}|\big] \; (u_a(\mathbf{x}) + C)$$

which is monotonically decreasing in both B and C. The utility of recourse is invariant w.r.t. B and C. Agents only perform recourse if manipulation yields lower utility gain, thus agent utility gain is monotone decreasing in B and C. A symmetric argument can be made for the principal's utility. □

Theorem 8. *When agent utility is constant, the expected number of agents who either choose to perform recourse or truthfully report is $nF_R\left(\min\left(\bar{u}_a, \frac{B(C+\bar{u}_a)}{n}\right)\right)$.*

Proof. This follows directly from Theorem 2. □

Lastly, we bound the fines paid by agents when the principal has budget B and the fine is C.

Theorem 9. *Let $F_R(k) = \mathbb{P}\big(c_R(\mathbf{x}, \mathbf{x}_R) \leq k\big)$ (CDF of c_R). Suppose agent utility is constant, define $C' \equiv C + \bar{u}_a$, and let A_M be the expected fines paid by agents. Then,*

$$BC\big(1 - F_R(2BC'/n)\big) \leq A_M \leq BC2\big(1 - F_R(BC'/n)\big)$$

Theorem 9 can be interpreted as quantifying the fines paid by the population in terms of how costly recourse is (i.e., the growth rate of F_R). If the principal audits B manipulating agents, the population pays $C \cdot B$. The terms $1 - F_R(2BC'/n)$ and $2(1 - F_R(BC'/n))$, in turn, approximate the probability that a given audit was conducted on a manipulating agent.

These bounds also express the parabolic nature of the fines paid by agents. For small B and C, the fines paid by agents are small (even if all agents manipulate). For large B and C, the cost of manipulation is sufficiently high that few agents will manipulate, and thus, average fines are small. It is the *intermediate* range of values of B and C for which both BC (fines paid when all audits are successful) and $1 - F_R(BC'/n)$ (probability of a successful audit) are large.

Proof Sketch of Theorem 9. The full proof is deferred to the Supplement, Section A.4. Given a set of reports $\mathbf{Z}'$, let n_M be the number of reports which are manipulations. Each report in $\mathbf{Z}'^{(1)}$ has equal probability of being audited, namely $B/|\mathbf{Z}'^{(1)}|$, implying that agents pay fines $\mathbb{E}\left[BCn_M/|\mathbf{Z}'^{(1)}|\right]$. We can bound $\mathbb{E}\left[n_M/|\mathbf{Z}'^{(1)}|\right]$ as

$$\mathbb{E}\big[n_M/n\big] \leq \mathbb{E}\big[n_M/|\mathbf{Z}'^{(1)}|\big] \leq \mathbb{E}\big[n_M/|\mathbf{X}^{(1)}|\big] \qquad (9)$$

The left-hand side follows from there being at least as many approved reports $|\mathbf{Z}'^{(1)}|$ as agents n. The right-hand side follows from $|\mathbf{Z}'^{(1)}|$ being greater than the number of approved truthful reports $|\mathbf{X}^{(1)}|$. Inequality 9 can be rewritten as,

$$1 - F_R\big(2BC'/n\big) \leq \mathbb{E}\big[n_M/|\mathbf{Z}'^{(1)}|\big] \leq 2\big(1 - F_R(BC'/n)\big)$$

Multiplying each side by BC completes the proof. □

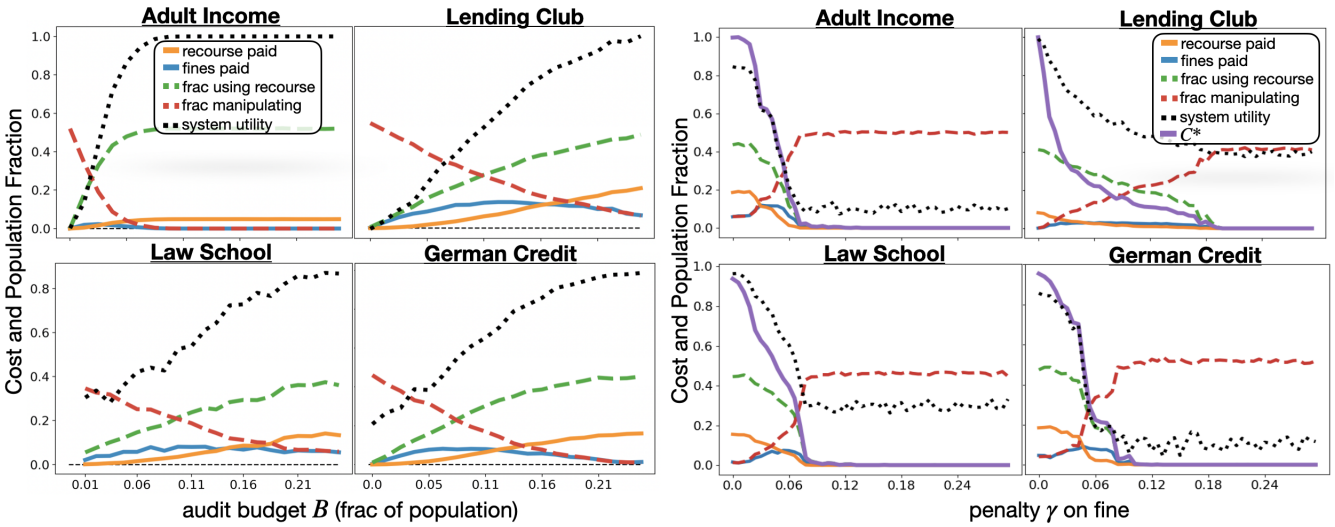

Figure 1: Fraction of agents choosing recourse or manipulation (green and red), average cost paid for each action (orange and blue), and system utility (black), for a fixed fine of $C = 1$ (left) or designed fines with audit budget $B = n/10$ (right). ' To estimate utility the principal uses Logistic Regression (top row) and 2-layer Neural Networks (bottom row).

8 Experiments

We conduct experiments using four common datasets: **Adult Income** [Kohavi and others, 1996], **Law School** [Wightman and Council, 1998], **German Credit** [Dua and Graff, 2019], and **Lending Club** [LendingClub, 2018], in which the objective is binary prediction. In the Adult Income and Law School datasets, 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 $y = 1$ and $u_p(\mathbf{x}) = -1$ when $y = 0$. In the German Credit and Lending Club datasets, agents have utility which is inversely proportional to their income and savings (credit is more valuable to those with lower existing capital); the principal's utility is equal to the total repayment of approved agents. The cost of recourse is $c_R(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|_2$. Full experimental details are provided in the supplement Section A.5.

We measure the fraction of the population performing recourse or manipulation, as well as the average cost incurred by agents for either action (Figure 1). In this figure three interesting phenomena occur. First, the average fines paid by agents is roughly parabolic in the audit budget B (Figure 1 left), and in the penalty γ which controls the size of the fine C^* (Figure 1 right). Thus, it is the intermediate values of B and C for which agents are most heavily fined. In these cases, B and C are not large enough to effectively dissuade manipulations, but are large enough to frequently catch and fine agents manipulating. This parabolic relationship is anticipated by Theorem 9. Second, the maximum cost spent on recourse exceeds the maximum fines paid. This is due to the fact that agents will only select recourse once the cost of manipulation is sufficiently high. Third, as the number of agents choosing recourse increases, so to does system utility. When an agent performs recourse, their true qualification improves (e.g., greater loan repayment), thus increasing the principal's utility when approving that agent.

Additionally, we measure the fraction of the audit budget which the principal allocates to subsides for varying subsidy functions (Figure 3 in Supplement Section A.5). As predicted by Theorem 4, we observe that when the subsidy function more effectively decrease recourse costs, both the allocation of subsides and the principal's utility increases. Thus, settings in which the cost of recourse is more easily offset give rise to a mutual benefit for both the system and individuals.

9 Conclusion

We investigated the relationship between manipulation and recourse when the principal possesses the ability to audit agent reports. We demonstrated that auditing can be used as an effective tool in preventing agent manipulation while still allowing the principal to offer recourse and maintain their desired classifier $\hat{y}$. For both a recourse-maximizing and utility-maximizing principal, the optimal audit policy is straightforward to execute, despite the seemingly complex nature of the problem. In particular, given a set of report $\mathbf{X}'$ the principal's best strategy is to uniformly audit all positively classified reports. Additionally we studied subsides, which allow the principal to allot a portion of their audit budget in order to decrease the cost of recourse. In this case, we find that when agent utility is constant, both objectives of recourse maximization and utility maximization are aligned; however, this is not the case for general agent utilities. Moreover, when agent utility is constant, the principal is guaranteed to spend a nonzero fraction of their audit budget on subsides, so long as the subsidy function $s(b)$ has better than linear scaling in b. Additionally we looked at the case when the principal posses the ability to select the fine for failing an audit, and again found that the objectives of recourse and utility maximization are aligned. Lastly we examined this problem from an empirical perspective and found that auditing can successful induce recourse as well as maximize system utility in practice.

Acknowledgements

This work was partially supported by the NSF (IIS- 1939677, IIS-1903207, IIS-1905558, IIS-2127752, IIS-2127754, IIS-2143895, and IIS-2040800), ARO (W911NF1810208), Amazon, and JP Morgan.

References

[Barsotti *et al.*, 2022] 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.

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

[Butaru *et al.*, 2016] Florentin Butaru, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W Lo, and Akhtar Siddique. Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72:218–239, 2016.

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

[Dong *et al.*, 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*, pages 55–70, 2018.

[Dua and Graff, 2019] D Dua and C Graff. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california, school of information and computer science. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[Estornell *et al.*, 2021] Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. Incentivizing truthfulness through audits in strategic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5347–5354, 2021.

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

[Haghtalab *et al.*, 2020] Nika Haghtalab, Nicole Immorlica, Brendan Lucier, and Jack Z. Wang. Maximizing welfare with incentive-aware evaluation mechanisms. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 160–166. International Joint Conferences on Artificial Intelligence Organization, 2020.

[Hansen, 1979] Eldon R Hansen. Global optimization using interval analysis: the one-dimensional case. *Journal of Optimization Theory and Applications*, 29(3):331–344, 1979.

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

[Karimi *et al.*, 2021] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362, 2021.

[Karimi *et al.*, 2022] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Comput. Surv.*, 55(5), dec 2022.

[Khandani *et al.*, 2010] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

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

[Kohavi and others, 1996] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[LendingClub, 2018] LendingClub. *Lending Club approved loans 2008-2018*. LSAC research report series. Lending Club, 2018.

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

[Locatelli, 1997] Marco Locatelli. Bayesian algorithms for one-dimensional global optimization. *Journal of Global Optimization*, 10(1):57–76, 1997.

[Lundy *et al.*, 2019] Taylor Lundy, Alexander Wei, Hu Fu, Scott Duke Kominers, and Kevin Leyton-Brown. Allocation for social good: Auditing mechanisms for utility maximization. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, page 785–803, 2019.

[Milli *et al.*, 2019] 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*, pages 230–239, 2019.

[Sergeyev, 1998] Yaroslav D Sergeyev. Global one-dimensional optimization using smooth auxiliary functions. *Mathematical Programming*, 81(1):127–146, 1998.

[Shavit *et al.*, 2020] Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal strategic linear regression. *International Conference on Machine Learning*, pages 8676–8686, 2020.

[Tsirtsis *et al.*, 2019] Stratis Tsirtsis, Behzad Tabibian, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf,

and Manuel Gomez-Rodriguez. Optimal Decision Making Under Strategic Behavior. *arXiv e-prints*, 2019.

[Upadhyay *et al.*, 2021] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16926–16937. Curran Associates, Inc., 2021.

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

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

[Wightman and Council, 1998] L.F. Wightman and Law School Admission Council. *LSAC National Longitudinal Bar Passage Study*. LSAC research report series. Law School Admission Council, 1998.