

Incentive-Aware Machine Learning; Robustness, Fairness, Improvement & Causality

CHARA PODIMATA
MIT

The article explores the emerging domain of incentive-aware machine learning (ML), which focuses on algorithmic decision-making in contexts where individuals can strategically modify their inputs to influence outcomes. It categorizes the research into three perspectives: *robustness*, aiming to design models resilient to “gaming”; *fairness*, analyzing the societal impacts of such systems; and *improvement/causality*, recognizing situations where strategic actions lead to genuine personal or societal improvement. The paper introduces a unified framework encapsulating models for these perspectives, including offline, online, and causal settings, and highlights key challenges such as differentiating between gaming and improvement and addressing heterogeneity among agents. By synthesizing findings from diverse works, we outline theoretical advancements and practical solutions for robust, fair, and causally-informed incentive-aware ML systems.

1. INTRODUCTION

Machine Learning (ML) algorithms are deeply embedded in various aspects of modern life, influencing everything from enhancing daily conveniences and shaping online purchasing behavior to making critical decisions in areas such as hiring, loan approvals, college admissions, and probation rulings. Given the high stakes of these decisions, individuals often have strong incentives to strategically modify the data they provide to these algorithms to secure more favorable outcomes. For instance, individuals might open additional credit accounts or take other steps to improve their credit scores before applying for a loan. In the context of college admissions, applicants may retake standardized tests like the GRE, enroll in test preparation courses, or even switch schools to boost their class rankings, all in efforts to present themselves as more competitive candidates.

Such instances of “strategic adaptation” have been extensively documented across disciplines including Economics, Computer Science, and Public Policy [Björkegren et al. 2020; Dee et al. 2019; Dranove et al. 2003; Greenstone et al. 2022; Gonzalez-Lira and Mobarak 2019; Chang et al. 2024]. The challenge arises when decision-makers deploying ML algorithms fail to account for these adaptations, potentially undermining the original goals of the policies the algorithms are intended to support. For example, in college admissions, a student’s decision to change schools solely to improve their class ranking may not necessarily reflect a substantive improvement in their qualifications.

It is important to note that not all strategic adaptations are inherently problematic. Some represent attempts to “game” the system (e.g., switching schools for a better ranking), while others involve genuine efforts at self-improvement (e.g., dedicating more time to study). The distinction between these types of adapta-

tions underscores the nuanced nature of this phenomenon and its implications for algorithmic decision-making.

What should decision-makers do when individuals are incentivized to alter the data they provide to ML algorithms in pursuit of better outcomes? And even if the learner manages to robustify (or calibrate) their algorithms to account for such behavior, what are the societal implications? These are some of the central questions addressed by the emerging field of *incentive-aware ML* (also known as “strategic classification” or “performative prediction”).¹

The purpose of this article is to provide an introduction to incentive-aware ML and an overview of the key results in the field. We categorize the literature on incentive-aware ML into three main perspectives: *robustness*, *fairness*, and *improvement & causality*. While some papers contain elements of multiple categories, we classify them based on their primary focus or central contribution. Broadly speaking, the “robustness” perspective adopts the viewpoint of the decision-maker, assuming that agents always attempt to “game” the decision rule. The goal in this context is to design algorithms that achieve optimality despite strategic adaptations by the agents. The “fairness” perspective examines the downstream societal impacts of algorithmic decision-making under varying assumptions about the agents’ capacity to strategically adapt. Lastly, the “improvement & causality” perspective recognizes that not all strategic adaptations are harmful; in some cases, agents’ adaptations in response to decision-making algorithms lead to genuine, fundamental improvements rather than merely fooling the algorithm. The distinctions among these perspectives, as well as the models and settings considered, will be formalized in the following section.

This article is organized as follows: Section 2 formalizes all the different formulations of the incentive-aware ML problem while giving some example reference papers for each modeling assumption; Section 3 presents a breakdown of the main contributions from the literature from the *robustness* perspective; Section 4 outlines the results that have been obtained from the *improvement & causality perspective*; Section 5 discusses the *fairness* perspective. Finally, Section 6 offers some parting thoughts on where the literature stands and where we should go next (according to the author’s personal opinions, at least).

2. OVERVIEW OF MODELS

In the problem of incentive-aware learning, there is an interaction between a *principal* (aka *learner*, *decision-maker*) and *agents*. The problem has been studied both in the *offline* (i.e., where there is one decision that is made by the principal and then the interaction stops) and *online* setting (i.e., where there are sequential decisions). Before we outline each setting, let us introduce some common notation.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ the *feature* space and $\mathcal{Y} = \{0, 1\}$ (resp. $\mathcal{Y} \subseteq [0, 1]$ for linear regression) the *label* (resp. response) space. We assume that the label (resp. response) is $y = h^*(x)$, where $h^* : \mathcal{X} \rightarrow \mathcal{Y}$ is called the *ground truth* function (which is not

¹Throughout this article, we use the terms “incentive-aware” and “strategic” ML interchangeably. The author prefers “incentive-aware” as it more comprehensively captures the considerations arising from agents’ behaviors. However, “strategic” is more commonly used in the literature.

necessarily linear).² We will denote by \mathcal{H} the *concept class* where h^* belongs to.

Let $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ be the principal’s loss function. Different applications of interest (within the general incentive-aware learning literature) call for different loss functions for the principal. Examples of frequently used loss functions for classification tasks include:

- (i) 0 – 1 loss (e.g., [Chen et al. 2020]): $\ell(y, y') := \mathbb{1} \{\text{sign}(y \cdot y') = 1\}$.³
- (ii) logistic loss (e.g., [Dong et al. 2018]): $\ell(y, y') := \log(1 + e^{-y \cdot y'})$
- (iii) hinge loss (e.g., [Dong et al. 2018]): $\ell(y, y') := \max\{0, 1 - y \cdot y'\}$

For the regression tasks, the most commonly used loss is some L_p norm.

As is the case in traditional ML, the choice of loss function for the principal affects what algorithms should be used, and what guarantees can be obtained.

Offline Setting

In the offline setting (e.g., [Hardt et al. 2016]), we assume that the agents’ features are drawn from some distribution \mathcal{D} . The interaction between the principal and the agent can be viewed as a *Stackelberg game* that plays out as follows:

- (1) Nature draws $x \sim \mathcal{D}$.
- (2) The principal —without knowing x — commits to (and publicly announces) a decision-making rule $f : \mathcal{X} \rightarrow \mathcal{Y}$.
- (3) The agents observe f and their point (x, y) .
- (4) Given f, x, y , the agents choose $\hat{x}(f)$ where $\hat{x}(f; x, y) \in \mathcal{X}$ is the best-response of the agent (given pair (x, y)) to the principal’s rule f .
- (5) The agent reports point $(\hat{x}(f; x, y), y)$ to the principal.

In Step (4), we are using $\hat{x}(f; x, y)$ abstractly; we are going to specify how it is computed later on. At a high level, $\hat{x}(f; x, y)$ is such that the agent obtains a better standing with regards to f (e.g., the agent gets classified as +1 from f in classification settings); see “Agents’ Response” for details. To simplify notation, we write $\hat{x}(f)$ (instead of $\hat{x}(f; x, y)$ when clear from context).

For now, let’s assume that when the agents best respond to a decision-making rule, they are *merely* trying to “game” it. We will contrast this approach to the Causality viewpoint, highlighted below.

In the “robustness” perspective, the principal’s goal is to find a function $f^* \in \mathcal{F}$ (where $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ is a hypothesis class over which we are searching) such that:

$$f^* = \arg \min_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mathcal{D}} [\ell(h^*(x), f(\hat{x}(f)))] \quad (1)$$

In words, in the “robustness” perspective for the offline learning setting, the principal’s goal is to find a function that minimizes the expected loss between the ground truth label (resp. response variable for regression) and the predicted label (resp. score) that function f assigns to the (potentially) altered datapoint $\hat{x}(f)$.

²Some works on incentive-aware linear regression assume that $y = h^*(x) + \varepsilon$ where ε is some small, zero mean noise, but that will not constitute a material difference in this article.

³We use $\text{sign}(x) = 1$ to denote that x is positive and $\text{sign}(x) = -1$ otherwise.

Online Setting

In the online setting (e.g., [Dong et al. 2018; Chen et al. 2020; Ahmadi et al. 2021]), the interaction between the principal and the agents happens repeatedly over T rounds. For every round $t \in [T]$, the interaction protocol is the following:

- (1) Nature chooses $x_t \in \mathcal{X}$.
- (2) The principal (without observing x_t) commits to (and publicly announces) decision-making rule $f_t \in \mathcal{F}$.
- (3) The agent observes f_t and their point (x_t, y_t) .
- (4) The agent chooses $\hat{x}_t(f_t; x_t, y_t)$ such that $\hat{x}_t(f_t; x_t, y_t)$ is the agent's best response (given pair (x_t, y_t)) to rule f_t .
- (5) The agent reports point $(\hat{x}_t(f_t; x_t, y_t), y_t)$ to the principal.

As we did for the offline case, for ease of notation, we will simply write $\hat{x}_t(f_t)$ in place of $\hat{x}_t(f_t; x_t, y_t)$ whenever clear from context. In the online setting, we assume that the principal knows the agents' utility function, but not the original point, x_t . The choice of $\hat{x}_t(f_t)$ in Step (4) depends on the agent's utility function; see "Agent's Response" below.

A couple of remarks are in order. First, the sequence $\{x_t\}_{t \in [T]}$ that the nature chooses can be *adversarial*. Second, \mathcal{F} can be a general class of functions. That said, the current literature only focuses on *linear* functions. Third, for the robustness perspective in online learning settings, we again assume that when the agent strategically adapts to a rule f_t , they can not influence their y_t (i.e., y_t remains the same both for x_t and for the misreport \hat{x}_t).

When we adopt the robustness perspective, the principal's goal is to minimize *Stackelberg* regret defined as follows:

$$\text{Reg}(T) := \sum_{t \in [T]} \ell(h^*(x_t), f_t(\hat{x}_t(f_t))) - \min_{f^{\text{OPT}} \in \mathcal{F}} \sum_{t \in [T]} \ell(h^*(x_t), f^{\text{OPT}}(\hat{x}_t(f^{\text{OPT}}))) \quad (2)$$

Note that similar to the offline model, we are comparing the algorithm's performance to the best fixed rule f^{OPT} had you given the agents the opportunity to best respond. In other words, we are comparing to the Stackelberg equilibrium rule.

Causality

So far, in both the offline and online settings, we have assumed that even after the agent strategically adapts, their y_t remains the same as it was prior to the adaptation (e.g., when an agent increases the number of credit cards they have, they have not actually improved their creditworthiness; they have merely tried to game the credit scoring system). This meant that *every* strategic adaptation was perceived as "gaming" and hence, the principal was trying to suppress it. However, for some applications of interest (e.g., for school admissions or loan approvals), some types of strategic adaptation are not gaming and should instead be encouraged or incentivized. For example, in a school admissions example, a strategic adaptation that makes the student study more in order to pass the threshold for admission is not gaming; rather, it is a way for the student to become a better potential candidate for the school of their choice.

To capture this, some settings in incentive-aware ML assume that in any d -dimensional feature vector, some features are *causal* (i.e., by changing them, the agent can change their actual y) while the rest are *proxy/non-causal* (i.e., by changing them, the agent cannot change their actual y). As a result, agent actions that change causal features have the ability to change the ground truth qualifications of an agent; as such, they can lead to *genuine improvement*, as opposed to the *gaming* which is induced by proxy features. The papers that assume causality of features (e.g., [Miller et al. 2020; Shavit et al. 2020; Bechavod et al. 2021]) use the language of *structural causal graphs* [Pearl 2009] in order to model the causal effects of the agents' different features.

Agents' Response

We next turn our attention to the way in which the agents choose their best response to the principal's algorithm. For an agent with ground truth feature vectors x , we use $u(x, \hat{x}; f)$ to denote the agent's *utility* for reporting \hat{x} when the principal uses classification/regression function f . We focus on utility functions of the form:

$$u(x, \hat{x}; f) := \text{val}(\hat{x}; f) - \text{cost}(x, \hat{x}) \quad (3)$$

where $\text{val}(\hat{x}; f)$ corresponds to the *value* that the agent obtains by reporting \hat{x} when the principal uses f , and $\text{cost}(x, \hat{x})$ corresponds to the *cost* that agent incurs for changing their feature from x to \hat{x} . There have been two types of value functions that have been primarily used in the literature:

- (i) (continuous) $\text{val}(\hat{x}; f) := f(\hat{x})$ (i.e., the value is just the evaluation of the function f for the reported feature \hat{x}) (e.g., [Dong et al. 2018; Bechavod et al. 2022; Shavit et al. 2020]).
- (ii) (discrete) $\text{val}(\hat{x}; f) := \gamma \cdot \mathbb{1}\{\text{sign}(f(\hat{x})) = 1\}$ (i.e., the agent cares only about being classified as passing a threshold (e.g., [Chen et al. 2020; Ahmadi et al. 2021])). Unless specified otherwise, we will use $\gamma = 1$.

As for the cost function, there have been primarily two families that the literature has considered:

- (i) (L_p -norm) $\text{cost}(x, \hat{x}) := \delta \cdot \|x - \hat{x}\|_p$ for some $\delta > 0$ (e.g., [Chen et al. 2020; Ahmadi et al. 2021; Bechavod et al. 2021]). The most frequently used norms are $p = 1$ and $p = 2$.
- (ii) (separable) $\text{cost}(x, \hat{x}) := c(\hat{x}) - c(x)$ (e.g., [Hardt et al. 2016; Hu et al. 2019]). These cost functions are suitable for settings where achieving each feature has a certain cost, but this is independent of which feature the agent started from.

The vast majority of the literature assumes that (given the aforementioned utilities) the agents are *best responding* to f , i.e., that $\hat{x}(f) = \arg \max_{x' \in \mathcal{X}} u(x, x'; f)$. There are some notable exceptions to this assumptions which we highlight in Section 3. Finally, most of the literature (except a few, e.g., [Dong et al. 2018]) assumes that *all* agents can respond strategically. For example, in classification, even agents with $y = 1$ may want to strategize, if they know that the classification rule f will classify them as 0.

Remark 2.1. In general, we assume that the principal knows the agents’ value and cost functions (including δ); they are only missing the original point x and can never fully learn it. To be more specific, given the value and cost functions, the reported \hat{x}_t and the y , the principal *cannot* reverse engineer the original x . There are a couple of works that focus on restricted strategic classification settings where δ is unknown, but the principal can still learn robust decision rules (see Section 3 for details).

Continuous Adaptation vs Manipulation Graph. Some works move away from the continuous⁴ model of strategic adaptation. Instead, they introduce the idea of a *manipulation graph* (e.g., [Ahmadi et al. 2023]). In incentive-aware learning with manipulation graphs, the assumption is that there exists a graph $G(\mathcal{X}, E)$ to capture all possible manipulations. In graph G each node corresponds to a different feature vector and each edge $e = (x, x') \in E$ captures the manipulation from x to x' . The cost function then $\text{cost}(x, x')$ is defined as the sum of costs to move from x to x' , if such a path exists in G . We will highlight which works use manipulation graphs instead of continuous adaptation in the coming sections.

Full vs Partial Information about the Principal’s Algorithm. We have so far assumed that the agent has *full* knowledge of f (or f_t) at the time of choosing their best response.⁵ Although this is a useful assumption to understand what solutions are possible in the worst case, in reality it is far from the truth; while agents do exhibit strategic adaptation, they seldom have *full* information about the decision-making rules used. There has been an emerging interest in modeling partial information from the agent side (e.g., [Braverman and Garg 2020; Ghalme et al. 2021; Bechavod et al. 2022; Cohen et al. 2024]), but no single model has prevailed as the canonical one. We highlight these models in the coming sections.

Heterogeneous Agents

Finally, we have so far assumed that there is a single \mathcal{D} representing the entire population and that every agent shares the same utility function. In other words, we have assumed that agents are *homogeneous*. However, this assumption is often unrealistic; for instance, in the context of school admissions, it is unlikely that everyone in the population has the same natural ability to succeed in school or the same capacity to take steps to improve their chances of being admitted.

Agent heterogeneity has been studied primarily in two different forms. First, agents may come from heterogeneous populations (i.e., their features and labels may originate from different distributions e.g., [Milli et al. 2019; Hu et al. 2019]). Second, agents may have different abilities to adapt to the decision rule that the principal is using (either because of different cost functions e.g., [Milli et al. 2019; Hu et al. 2019]) or because of different understanding of the decision rule (in the case of partial information) (e.g., [Bechavod et al. 2022]). We discuss heterogeneous agents in Section 5.

⁴The literature sometimes refers to this type of strategizing as “ball manipulation”.

⁵Historically, this is a byproduct of the fact that the original papers modeled the paper as a Stackelberg game. In the Stackelberg games literature, the standard assumption is that the principal announces their strategy at the beginning of the interaction with the agent. This announcement gives them “commitment power” (as it is referred to in that literature).

Remark 2.2. As should be clear by now, this article focuses exclusively on strategic adaptation that occurs in the *feature* space, rather than the *label* or *response variable* space. There have also been a series of works on ML algorithms when the agents can strategically adapt their label (e.g., [Dekel et al. 2010; Chen et al. 2018]) but they are beyond the scope of this article. The aforementioned articles take a “robustness” perspective.

One final note: the terminology introduced in this section will be used throughout the following sections to describe each paper. This consistent terminology is intended to help the reader develop a clear mental framework for understanding the types of results obtained for each model variant of incentive-aware learning.

3. ROBUSTNESS PERSPECTIVE MAIN RESULTS

We begin our exposition with the *robustness* perspective. In this framework, the principal seeks to learn the most accurate decision-making rule (as defined in Equation (1)) that maps agent features to a score or classification label, thereby minimizing their loss. Simultaneously, agents strategically manipulate the data they submit to the decision-making rule in an effort to “game” the system. We first examine the offline/batch and online learning settings in Sections 3.1 and 3.2, respectively, focusing on scenarios where agents have full knowledge of the principal’s decision-making rule. Subsequently, in Section 3.3, we explore settings where agents have only *partial* information about the principal’s decision-making rule. Finally, we conclude this section by discussing cases where agents are *not* individually rational when selecting their misreports, $\hat{x}(f)$, in Section 3.4.

3.1 Offline and Batch Learning Setting

[Hardt et al. 2016] introduced the problem of “strategic classification” in the *offline* setting and formulated it as a Stackelberg game. In their framework, the population of agents is assumed to be *homogeneous*, with each agent aiming to maximize their probability of being classified as +1 while incurring a cost for doing so. The principal, on the other hand, wants to design a classifier that converges to the offline optimal in terms of “accuracy” (as defined in Equation (1)) for the 0 – 1 loss. The agents are assumed to have *full* information about the classification rule and are best-responding to it. The authors show that for agents with separable cost functions, it is possible to design efficient and nearly optimal classifiers, even for concept classes that are computationally hard to learn. Their theoretical framework further includes impossibility results for learnability when the agents have *general* cost functions, illustrating the fundamental challenges of achieving classification robustness against strategic behavior.

Working in the *offline* or *batch* setting with a *homogeneous* population of agents, [Levanon and Rosenfeld 2021] introduce the notion of *strategic empirical risk minimization* (strategic ERM) as an approach for designing strategy-robust decision rules for the principal. At a high level, the authors propose a “smoothed” version of the strategic classification problem, incorporating the agents’ best-response behavior as a function of the decision rule f into the optimization process for f . While the paper does not provide theoretical guarantees, it includes a series of experiments demonstrating how strategic ERM might perform in practice. However,

the assumption of a “smoothed” version of the problem has limitations from a real-world modeling perspective. As noted by several works (e.g., [Dong et al. 2018; Chen et al. 2020]), the motivating settings for strategic classification often make it infeasible to identify a “smooth” loss function for the principal once the agents’ best-response behavior is incorporated.

Still working within the ERM paradigm, in the *offline* learning setting and drawing intuition from traditional PAC learning [Valiant 1984], there has also been interest in a PAC version of incentive-aware learning, i.e., given a set of points that have been strategically modified, identify the complexity of finding a classification function that is $\varepsilon(\eta)$ -optimal (according to Equation (1)) with high probability at least $1 - \eta$. This version of the problem was introduced by [Zhang and Conitzer 2021]. The authors assume that the agents can best-respond according to a *reporting structure* which maps original features to manipulated ones.⁶ Moreover, they assume that the principal is facing a *homogeneous* population of agents. The paper first shows that the vanilla ERM (i.e., the one ignoring incentives) has poor performance in strategic settings.⁷ Subsequently, they show that a version of *strategic empirical loss* can obtain nearly optimal sample complexity bounds. To construct their strategic empirical loss, the authors take a “worst-case perspective”; for each reported point, they substitute it with the worst-possible original point it could have originated from.⁸

In a similar vein, [Lechner and Uerner 2022] study the learnability of general concept classes with a new class of loss functions called *strategic loss* (which is used as a proxy hypothesis class for the principal). In their setting, the agents can manipulate according to a manipulation graph. The strategic loss is a discrete loss function which takes a value of 1 every time that either $f(x) \neq y$ (i.e., incorrect classification) or $f(x) = 0$ but there exists a point x' such that x' is a reachable misreport from x and $f(x') = 1$, and 0 otherwise. This new loss function aims to not only account for accuracy but also, for the societal burden that is induced when the agents fool the classifier.

[Sundaram et al. 2023] take incentive-aware PAC learnability one step further; the agent population is now *heterogeneous* (i.e., the cost function is the same across agents, but each agent may have a different γ in their value function), the principal does *not* know the agents’ value functions, but the principal has access to a training dataset that is un-manipulated (i.e., the principal can see some original x ’s). The key contribution of the work is the introduction of the *Strategic VC-Dimension* (the strategic analogue of VC-dimension), which quantifies learnability in settings where test data is strategically manipulated based on *heterogeneous* agents. The authors subsequently characterize the statistical and computational limits of strategic linear classification. This study also explores the role of *randomization* in improving accuracy under strategic manipulation. We expand on the role of *randomness* in

⁶This can be considered as part of the general “manipulation graph”-type of cost functions.

⁷For the online setting, a slightly stronger result of two-way incompatibility between regular and strategic settings was obtained by [Chen et al. 2020]. Specifically, the authors show that there exist classification settings for which every no-external regret algorithm incurs linear Stackelberg regret and vice versa.

⁸A version of this technique was also used in [Chen et al. 2020], albeit for the online version of the problem.

strategic classification settings in Section 3.3.

[Rosenfeld and Rosenfeld 2024] focus on learning a *linear* classifier (the principal has access to a set of un-manipulated data at training time), the agents have a 0 – 1 value function, and L_p -norm cost function. Importantly, the authors assume that δ (i.e., the cost function) is *not* known by the principal but is the same⁹ across all agents; yet, the principal still needs to learn a classification rule that converges to the optimal one. The authors take a robust optimization approach, by minimizing the worst-case risk over a family of costs which includes the target (unknown) cost. They do so, because as they show, if the principal has to commit to a single fixed cost for their risk minimization problem, then ERM can never provide a non-trivial data-independent guarantee (unless the assumed single fixed cost was precisely correct). As for the ERM, the authors consider a type of *hinge* loss, that is appropriately expanded in order to include the uncertainty induced by the unknown cost function. The main result of the paper is an efficient iterative algorithm that converges to the minimax optimal solution with rate $\tilde{O}(1/\sqrt{T})$, where $\tilde{O}(\cdot)$ hides polylogarithmic terms, and T is the number of the algorithm’s iterations.

3.2 Online Learning Setting

The online learning version of strategic classification was first studied by [Dong et al. 2018]. In their paper, the authors provide linear strategic classification algorithms with sublinear Stackelberg regret (see Equation (2) against a *homogeneous* population of agents with *linear* values (i.e., the agents care about maximizing their distance from the classifier, while being labeled as +1 by it). To give an overview of their approach, let w_t be the normal vector corresponding to classifier f_t for each round $t \in [T]$, i.e., $f_t(x) := w_t^\top x$. The main result of the paper is to find the sufficient conditions on the agents’ $\hat{x}(w_t)$ such that $\ell(w_t, \hat{x}(w_t))$ is *convex* in w_t , when $\ell(w_t, \hat{x}(w_t))$ is either the hinge or logistic loss. This task boils down to identifying the sufficient conditions on the agents’ cost function in order for $\ell(w_t, \hat{x}(w_t))$ to be convex in w_t . Convexity is desired, since if $\ell(w_t, \hat{x}(w_t))$ is convex in w_t , then the principal can apply any off-the-shelf bandit convex optimization algorithm and obtain sublinear Stackelberg regret. The paper obtains improved regret bounds under the assumptions that all agents with $y_t = 1$ are non-strategic.

But what happens when $\ell(w_t, \hat{x}(w_t))$ is not a convex function of w_t ? In an effort to answer this question in a general way, [Chen et al. 2020] studied online learning of linear classifiers in the following setting: the agents have a discrete value for passing the classifier (i.e., they obtain a value of 1 for passing the classifier and 0 otherwise), their cost function is δ -bounded (i.e., $\|\hat{x}_t(f_t) - x_t\| \leq \delta, \forall t \in [T]$), and the learner cares about the 0 – 1 loss. Importantly, the results of the paper do not require the agents to *rationally* best-respond; instead, knowing that the $\hat{x}_t(f_t)$ satisfy the constraint that $\|\hat{x}_t(f_t) - x_t\|_2 \leq \delta$ is enough. The paper provides a nearly tight algorithm that dynamically and adaptively partitions the space of feasible classifiers for the principal as new agents arrive. The final Stackelberg regret bound depends on the *instance* of datapoints $\{(x_t, y_t)\}_{t \in [T]}$ that nature chooses. The key tricks that the authors use is that when the principal sees a reported point $\hat{x}_t(f_t)$, then they know for sure that the true x_t lies inside a ball B , where

⁹This is the main difference with the *model* of [Sundaram et al. 2023].

$B := \{x \in \mathcal{X} : \|x - \hat{x}_t(f_t)\|_2 \leq \delta\}$. The final trick is to observe that given this information and the fact that the learner cares about the 0 – 1 loss, then the principal can obtain perfect information about the loss that would have been incurred in that round t if the same agent at round t were to best respond to some other normal vectors w for which $\|\hat{x}_t(w) - \hat{x}_t(f_t)\|_2 \leq 2\delta$. The theoretical analysis of the algorithm requires knowing the magnitude of the agents’ manipulation (δ) and access to a carefully crafted oracle that can provide some extra information to the principal about the structure of the agents’ unmanipulated data.

The aforementioned paper trades efficiency for generality. When the sequence of data $\{(x_t, y_t)\}_{t \in [T]}$ chosen by nature is *separable* by a margin, [Ahmadi et al. 2021] introduce a variant of the Perceptron algorithm, called the *Strategic Perceptron*, which is *computationally efficient* and converges to a maximum-margin classifier while making a bounded number of mistakes. The upper bound on the number of mistakes depends on the margin of the original, unmanipulated data and the agents’ strategizing power. The Strategic Perceptron is analyzed under the assumption that agents incur either L_1 or L_2 costs when misreporting from x to \hat{x} , and are rationally best-responding. Notably, the paper shows how to leverage the structure of the agents’ utility function together with the fact that the agents are rationally best responding to establish bounded mistake guarantees *even when* the magnitude of the manipulation cost is *not* known to the principal a priori — a result that was not achievable in [Chen et al. 2020].

Next, we transition from models of continuous strategic adaptation to models where agents determine their \hat{x}_t based on a manipulation graph, highlighting the work of [Ahmadi et al. 2023]. This setting generalizes the frameworks of [Zhang and Conitzer 2021] and [Lechner and Uerner 2022] to the online setting. The paper demonstrates that, unlike in the non-strategic classification setting, the vanilla Halving algorithm may incur an infinite number of mistakes. To address this, the authors propose a general algorithm for the strategic setting with a mistake bound of $O(\Delta \ln(|\mathcal{H}|))$, where Δ is the degree of the manipulation graph and \mathcal{H} is the (known) class of the target function. Furthermore, the paper extends the algorithm to the agnostic learning setting.

Adopting a similar perspective of testing the limits of strategic learnability, [Cohen et al. 2024] and [Ahmadi et al. 2021] investigate whether the learnability of a concept class implies its strategic learnability. They essentially show that every learnable function class remains learnable even when data is strategically manipulated. Both works model the agents’ feasible manipulations using manipulation graphs and consider scenarios where the graph is either fully known or only partially known to the principal. [Ahmadi et al. 2021] introduce the “strategic Littlestone dimension,” which captures the complexity of the agents’ manipulation graph and the hypothesis class. Both papers analyze strategic learnability across multiple variations of the baseline strategic classification model. Finally, [Shao et al. 2024] study learnability in terms of mistake bounds and sample complexity when agents’ manipulations are *heterogeneous*. They consider both continuous adaptations and manipulation graphs. As for the principal, they assume that some knowledge of x_t is available either before choosing the classification rule f_t or immediately afterward.

In a slightly different setup, [Harris et al. 2023] consider an online setting where

at each round the principal commits to a function $f_t : \mathcal{X} \rightarrow \{0, 1\}$, the agents can strategically adapt within a ball of radius δ of their true datapoint x_t , and the reward that the principal receives is linear in the agent’s unmodified context; more concretely, for each decision $\alpha \in \{0, 1\}$ the reward of the principal for a context x_t is: $r_t(\alpha) = \theta_\alpha^\top x_t + \varepsilon$, where θ_α is a d -dimensional vector. The authors assume that the principal has “apple tasting” feedback, i.e., the principal can observe $r_t(\alpha)$ only when $\alpha = 1$ (which in turn, is decided by the function f_t). The authors present algorithms that actually incentivize agents to be *truthful* (i.e., report x_t without any manipulation) while achieving sublinear regret.

3.3 Partial Information about the Principal’s Algorithm

So far, we have primarily assumed that the principal commits to a *deterministic* rule and that agents fully observe this rule. [Braverman and Garg 2020] were the first to highlight the role of *randomness* in the principal’s classifier and the impact of *noise* in the agents’ features on the outcomes of the strategic classification game. The paper demonstrates that to maximize accuracy (as defined by Equation (1)), the principal may *need* to employ randomized rules. This result creates an intriguing policy dilemma: on the one hand, randomized rules may be necessary to achieve optimal accuracy; on the other hand, their deployment can be legally problematic. Interestingly, the paper shows that introducing (or having inherently) noisier signals for the agents’ features can improve both accuracy and fairness in equilibrium across different subpopulations.

[Ahmadi et al. 2023] also explore the role of randomness in strategic classification, focusing on its impact on learnability. They consider two sources of randomness. In the first, the principal commits to a probability distribution over classifiers, thereby inducing certain probabilities of classification as +1 for agents. In the second, the principal commits to a probability distribution over classifiers, nature (which may adversarially select the next x_t) responds to this distribution, and the chosen agent x_t best responds to the *realized* classifier. The second model is more *transparent* to the agents than the first and enables the principal to design algorithms with improved regret guarantees.

If the principal has the choice between a transparent and an “opaque” classifier, which approach minimizes prediction error? [Ghalme et al. 2021] address this question in the setting of [Hardt et al. 2016] (i.e., offline, homogeneous population of agents, etc.). They define the *price of opacity* as the difference in prediction error when agents respond to a fully transparent classifier f versus an opaque rule \hat{f} . The paper studies the conditions under which the price of opacity can be positive or negative. Consistent with the theory of Stackelberg games, revealing f (or allowing it to be fully anticipated or deduced from \hat{f}) can sometimes benefit the principal, as it enables them to precisely predict how agents will react.

[Cohen et al. 2024] introduce a Bayesian classification setting, where the principal gradually reveals information about the classification rule. In this model, agents share a common distributional prior over the classifier used by the principal and best respond by maximizing their expected utility. The principal, in turn, can strategically release partial information about the classifier over time. The authors show how to release this information carefully to ensure that truly qualified agents

(i.e., $y_t = +1$) can pass the classifier while preventing unqualified agents from gaining sufficient information to successfully strategize and game the system.

Finally, [Bechavod et al. 2022] study a setting where agents acquire information about the classifier through “peer learning.” The primary focus of this work is on the fairness implications of information discrepancies across different subpopulations. Therefore, we defer a detailed discussion of this work to Section 5.

3.4 Beyond Rational Best-Response Agents

So far, we have focused on settings where agents best-respond to the principal’s rule. We now shift our attention to scenarios where agents do *not* precisely best-respond.

Although the results in [Chen et al. 2020] hold for *any* agent manipulation within δ of the true data point, [Jagadeesan et al. 2021] formalize alternative models for agent behavior that deviate from exact, rational best response. The authors demonstrate the brittleness of standard strategic classification algorithms when agents do not strictly adhere to the assumed best-response model. To address this, they identify a set of desiderata for agent responses that ensure algorithm stability and propose the *noisy response* model. In this model, agents best respond to a noise-perturbed version of the decision rule, inspired by the principles of smoothed analysis [Spielman and Teng 2009].

[Ebrahimi et al. 2024] study the role of behavioral biases in agents’ responses within strategic classification settings. Specifically, they consider agents who, when evaluating the value of passing the classifier, *weigh* the classifier’s features according to their own biases. The paper analyzes a homogeneous population of agents who can incur a cost of up to B for misreporting. It identifies cases where agents overshoot or undershoot the classifier’s boundary due to their biased perceptions of the classifier’s feature weights.

[Lechner et al. 2023] examine settings where the principal faces two sources of uncertainty regarding the agents’ responses. First, agents are not required to rationally best-respond and are instead permitted to use any *feasible* response that enables them to fool the classifier. Second, the principal does not have full knowledge of the agents’ manipulation graph but only knows the general family to which it belongs. Focusing on strategic loss, the authors study the learnability of both proper and improper learning under these assumptions. Their key result is that it is possible to learn an almost-optimal classifier in terms of strategic loss, even without precise knowledge of the manipulation graph.

[Cohen et al. 2024] explore the effects of partial knowledge of the manipulation graph on learnability. They show that when the principal knows only the general family of graphs to which the manipulation graph belongs, they can achieve nearly tight bounds on both sample complexity and regret. Furthermore, the difference in learning complexity between the fully-known and partially-known graph settings is (roughly) logarithmic in the size of the graph family.

Finally, [Ahmadi et al. 2024] also assume that the principal knows only the *family* of graphs to which the agents’ manipulation graph belongs. They derive a regret bound that is approximately optimal for certain instances. Additionally, they extend their results to a setting where each agent may have a different manipulation graph, provided all graphs belong to the same family. This generalized setting is referred to as the “agnostic” case.

4. IMPROVEMENT & CAUSALITY PERSPECTIVE MAIN RESULTS

Oftentimes strategic adaptation to algorithmic decision-making rules may lead to genuine improvement for the individuals; for instance, paying-off prior debt as a means of increasing your credit score actually helps you become more creditworthy. To state this in the language of incentive-aware learning, this means that the agents’ label y can change when they switch from their true x to the strategically manipulated \hat{x} . This section focuses on settings where strategic adaptation can lead to both gaming and actual improvement.

4.1 Improvement & Recourse

According to the “improvement”/“recourse”¹⁰ perspective, any strategic adaptation results in genuine improvement for individuals; that is, when a data point changes from x to \hat{x} , it holds that $h^*(x) < h^*(\hat{x})$.

[Kleinberg and Raghavan 2020] introduce one such model where agent “manipulations” result in changes to the underlying features, which can constitute genuine improvement for the agents. Their primary result shows that simple linear mechanisms suffice to incentivize genuine improvement in settings where the principal interacts with a single agent. [Harris et al. 2021] extend the model of [Kleinberg and Raghavan 2020] to settings where agents achieve improvements over a sequence of rounds, i.e., agents transition through different states over time as they respond to the principal’s rule.

[Alon et al. 2020] generalize the single-agent setting of [Kleinberg and Raghavan 2020] to a multi-agent framework. In their model, all agents share the same initial feature representation, but their ability to manipulate (quantified by their manipulation costs) differs.

[Haghtalab et al. 2020] also study multi-agent settings, focusing on designing evaluation mechanisms that maximize population-wide quality scores when agents can strategically alter their features at a cost. Their model differs from [Alon et al. 2020] in that agents can have different initial feature representations. The authors analyze two specific classes of mechanisms: linear mechanisms and linear threshold mechanisms. For linear mechanisms, they show that the optimal strategy corresponds to projecting the true quality function onto the observable feature space, which is computationally efficient. For linear threshold mechanisms, they develop approximation algorithms, including a constant-factor approximation algorithm under smooth feature distributions, that balance the trade-offs between incentivizing improvements and maximizing welfare. The paper further considers scenarios where the feature distribution is unknown and provides sample-complexity guarantees for learning optimal mechanisms.

[Tsirtsis and Gomez Rodriguez 2020] explore the design of optimal decision-making policies and counterfactual explanations in incentive-aware learning. They

¹⁰The term “recourse” comes from the traditional ML literature. Loosely speaking, algorithmic recourse refers to the ability of individuals to reverse negative decisions made by algorithms through counterfactual explanations provided alongside the decision. A substantial body of work exists on algorithmic recourse (see, e.g., [Karimi et al. 2020] for a survey), but it is beyond the scope of this article. Here, we focus specifically on the effects of strategic adaptation on algorithmic recourse.

model this problem as a Stackelberg game, where decision-makers provide *counterfactual explanations*—guidelines on how agents can change their features—and agents respond strategically to maximize their benefit. Unlike the standard Stackelberg game for incentive-aware learning, where the decision rule is announced, here the principal announces counterfactual explanations. The authors show that optimizing the set of counterfactual explanations for a fixed decision policy is NP-hard but can be addressed using approximation algorithms that leverage the problem’s submodularity. They further extend the problem to jointly optimize both the decision policy and explanations, reducing it to a non-monotone submodular maximization problem solvable with approximation guarantees. Additionally, the paper incorporates diversity constraints to ensure equitable distribution of explanations across populations.

Finally, [Bechavod et al. 2022] study the “improvement” perspective when the principal’s decision rule is not fully known to the agents. Their work focuses on the effects of information discrepancies across different subpopulations and is therefore discussed in Section 5.

4.2 Causality

How can we distinguish between agent actions that lead to genuine improvement versus those that constitute mere gaming? As [Miller et al. 2020] observe, designing “good” incentive-aware decision-making rules—rules that incentivize actions leading to genuine improvement while disincentivizing gaming—is equivalent to identifying the causal model underlying the setting (i.e., performing causal inference). Their work was the first to formalize this connection, introducing causal graphs to study how certain agent features affect (or do not affect) the target variable y .

Building on the theme of causality in incentive-aware learning, [Shavit et al. 2020] study incentive-aware linear regression, where the decision-maker seeks to optimize one of three objectives: (1) Agent Outcome Maximization (incentivizing agents to improve their outcomes), (2) Prediction Risk Minimization (ensuring accurate prediction of post-gaming outcomes), and (3) Parameter Estimation (accurately estimating the causal parameters of the outcome-generating process). The authors propose efficient algorithms for each objective in a realizable linear setting, leveraging the ability to test and observe agent responses to decision rules—effectively performing causal interventions. This ability to perform interventions makes their setting more tractable compared to [Miller et al. 2020]. Additionally, they address challenges such as omitted variable bias and interactions between observed and hidden features, which can undermine naive regression approaches. Extending beyond linear regression, [Horowitz and Rosenfeld 2023] study (agnostic) incentive-aware classification under causality with the goal of improving the principal’s accuracy.

In concurrent and independent work, [Bechavod et al. 2021] explore incentive-aware linear regression and demonstrate how agents’ strategic behavior can facilitate the learning of causal variables. The authors propose a batch-based retraining approach that iteratively updates the regression model, leveraging agents’ strategic modifications to improve predictive accuracy while incentivizing genuine improvement in features. They prove that this dynamic interaction enables the principal to accurately recover the true regression parameters over time, even when features are correlated. As a result, the principal can both incentivize genuine improvement

and improve the robustness of the decision model.

Finally, [Ahmadi et al. 2022] study incentive-aware classification under a causal model, addressing both discrete strategic adaptation (via manipulation graphs) and continuous adaptation. For the general discrete model, the authors design efficient algorithms to maximize true positives while ensuring no false positives, thus guaranteeing that only genuinely qualified agents are classified positively. They further show that the problem of selecting criteria to maximize true positives while allowing even a bounded number of false positives becomes NP-hard. In the continuous adaptation (linear) model, they develop algorithms to determine whether a linear classifier exists that classifies all agents accurately while incentivizing all improvable agents to become qualified.

4.3 Performativity

Before we conclude the section on improvement and causality in incentive-aware ML settings, we briefly touch on the literature on *performative prediction* [Perdomo et al. 2020]. Performative prediction is another framework to explain and reason about how predictions, when used to inform decisions, influence the outcomes they aim to predict. The authors develop a risk minimization framework and propose a new equilibrium notion called performative stability. Roughly speaking, this notion ensures that predictions are calibrated not to past data but to the outcomes they induce. The paper presents necessary and sufficient conditions for retraining algorithms to converge to performatively stable solutions with near-minimal loss. The main distinction between performative prediction and the other models that we highlight in this survey is that performative prediction uses certain smoothness assumptions on the way that original points x leads to strategically adapted points \hat{x} , instead of focusing on the agents' utility functions.

5. FAIRNESS MAIN RESULTS

Most (if not all) of the papers discussed so far in this article have focused on a homogeneous population of agents with which the principal is interacting. However, when the principal is interacting with a *heterogeneous* population of agents, with (potentially) different abilities to strategize and different qualifications, then optimizing for the desiderata of robustness-to-gaming or accuracy may have disparate downstream effects to the different subpopulations.

[Hu et al. 2019] and [Milli et al. 2019] independently and concurrently initiated the study of the disparate downstream effects of designing robust-to-gaming classifiers to different subpopulations. [Milli et al. 2019] defined the *social burden* of a classifier as the aggregate of the minimum cost an individual needs in order to be classified as a +1. For example, for agents with $y_t = +1$, a high social burden means that it is very costly for the agents to obtain their correct classification. The authors prove a general trade-off between principal's accuracy and agent utility. They also prove that when agents incur cost as a consequence of a principal making their classifier robust to strategic behavior, the costs can disproportionately fall on the disadvantaged subpopulations.

In a similar theme, [Hu et al. 2019] study negative externalities of strategic classification, and show that the Stackelberg equilibrium classifier leads to only false negative errors on the disadvantaged subpopulation but only false positives on the

advantaged population. Not only that, but they also show that providing a cost subsidy (whose goal is to counterbalance this the difference in false negatives and false positives from each subpopulation) can *actually* lead to worse outcomes for *everyone* in the game.

Focusing on the goal of group fairness, [Estornell et al. 2023] explores the unintended consequences of using fairness-aware algorithms in environments where agents can strategically manipulate their features to achieve better outcomes. While fairness in algorithmic decision-making is typically aimed at ensuring equitable treatment across demographic groups, the paper identifies a phenomenon called “fairness reversal”. This occurs when a fairness-driven classifier (designed to equalize outcomes between groups) becomes less fair than a conventional accuracy-focused classifier due to strategic feature manipulation by agents. The authors empirically demonstrate this phenomenon using benchmark datasets and attribute it to the selectivity of fair classifiers, which achieve fairness by excluding individuals from the advantaged group rather than including more from the disadvantaged group. They prove that increased selectivity is a sufficient, and in some cases necessary, condition for fairness reversal. They further show that fairness reversal does not occur when fairness is achieved through inclusiveness, where the classifier broadens access to the disadvantaged group.

The focus of the aforementioned works was on fairness in terms of classification accuracy. Lately, some works have started considering fairness in terms of improvement or recourse ability. [Gupta et al. 2019] address fairness in terms of *recourse*, i.e., the effort required to reverse a negative classification, across demographic groups. Mathematically, recourse is measured as the distance from an individual’s features to the decision boundary of a classifier. The paper introduces a new approach to regularize classifiers, minimizing disparities in recourse while maintaining predictive accuracy. It extends prior work on linear classifiers to more complex settings, including non-linear models and model-agnostic scenarios, where the decision boundary is not explicitly known. For the model-agnostic setting, the paper assumes that the agents have black-box access to the classifier, rather than the full mathematical formulation.

[Bechavod et al. 2022] study how disparities in information about decision rules affect the ability of agents from different sub-populations to improve their outcomes in strategic learning contexts. Unlike most traditional models that assume agents fully know decision rules, this work focuses on scenarios where decision rules are not fully known originally, and agents infer them based on their peers’ experiences, creating group-specific information levels; they refer to this process as “peer learning”. The study reveals that even when decision rules are optimized to maximize welfare, disparities in information and effort costs can lead to some sub-populations experiencing a decline in quality (“negative externality”). However, under specific conditions (e.g., proportional costs across groups or minimal information overlap — measured through an “information overlap proxy” metric — across groups) negative impacts can be mitigated.

[Ahmadi et al. 2023] study the problem of designing short-term goal structures to incentivize agents with varying abilities to improve their skills or capacities-for-improvement. It proposes two models: (1) the common improvement capacity

model, where all agents share the same improvement limit, and (2) the individualized improvement capacity model, where agents have personalized improvement limits. The authors develop algorithms to optimize the placement of target skill levels (i.e., goals) to maximize social welfare (i.e., total improvement across all agents) and ensure fairness among groups. One challenge they address is the non-monotonic nature of social welfare, where adding new target levels may unintentionally reduce overall improvement. Finally, they present an extension for the case where the principal has sample access to the available data when designing the classifier.

6. CONCLUSION

The purpose of this article has been to provide a gentle introduction to the exciting area of incentive-aware ML. We categorized the existing research into *robustness*, *fairness*, and *improvement/causality perspectives*, and we highlighted the diverse approaches and objectives within each domain. We outlined some of the foundational models and theoretical frameworks for understanding strategic adaptation, from offline and online learning settings to causal perspectives, and we emphasized the complexities introduced by agent heterogeneity and partial information.

There have also been a handful of topics related to incentive-aware ML settings that we did not touch upon, as they did not directly fit under one of our three outlined perspectives. Examples include: [Zrníc et al. 2021] who study how the Stackelberg game (and its outcomes) change when the principal and the agent (termed “leader” and “follower” in their paper) alternate in order; papers on econometrics for strategic agents (e.g., [Harris et al. 2022; Harris et al. 2024]); and papers focusing on agents that can choose to not participate in the algorithmic decision making process, if that is aligned with their utility maximization (e.g., [Krishnaswamy et al. 2021], [Horowitz et al. 2024]).

For all the excitement surrounding this research area, there is one question that seems as pressing as ever.

What comes next for the literature on incentive-aware ML?

In the author’s view, there are two primary paths for the future of incentive-aware ML. The first path is the more well-established and widely explored. There remain myriad settings requiring theoretical analysis of the interactions between individuals and a decision-making principal. For example, how do information discrepancies about the principal’s algorithm across different subpopulations affect their abilities to genuinely improve their outcomes versus merely game the system? Are there properties of “interpretable” decision-making algorithms that can provably incentivize genuine improvement rather than gaming? Developing new models and providing provable guarantees for these questions will help solidify the theoretical foundations of incentive-aware ML.

The second path is less charted and relatively unexplored, particularly from a theorist’s perspective. Although examples of individuals strategizing and adapting to algorithmic decision-making rules are abundant, incentive-aware ML still needs to identify a *concrete* application domain where the insights gained from theoretical advancements can *actually be applied*. Such a domain would allow incentive-aware algorithms to be deployed and evaluated against other “robust” algorithms.

This approach differs from the path the literature has predominantly taken. To illustrate this distinction, consider the steps required to apply theoretical insights from incentive-aware ML to a practical domain, such as recommendation systems (RecSys).¹¹ To apply these insights effectively in the RecSys domain, we would need to address several questions: Do users “strategize” with their data (see e.g., [Haupt et al. 2023])? What utility function are they optimizing for? What does it mean for users to have “partial” information about the RecSys? What specific interventions can the RecSys implement to mitigate inequalities between different user subpopulations?

Identifying such a concrete application domain would enable the foundational results in this field to be translated into actionable insights, driving meaningful, real-world change. The author is optimistic about the potential of the next generation of incentive-aware ML research to bridge this gap and create significant societal impact.

REFERENCES

- AHMADI, S., BEYHAGHI, H., BLUM, A., AND NAGGITA, K. 2021. The strategic perceptron. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 6–25.
- AHMADI, S., BEYHAGHI, H., BLUM, A., AND NAGGITA, K. 2022. On classification of strategic agents who can both game and improve. In *3rd Symposium on Foundations of Responsible Computing, FORC 2022, June 6-8, 2022, Cambridge, MA, USA*, L. E. Celis, Ed. LIPIcs, vol. 218. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 3:1–3:22.
- AHMADI, S., BEYHAGHI, H., BLUM, A., AND NAGGITA, K. 2023. Setting fair incentives to maximize improvement. In *4th Symposium on Foundations of Responsible Computing, FORC 2023, June 7-9, 2023, Stanford University, California, USA*, K. Talwar, Ed. LIPIcs, vol. 256. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 5:1–5:22.
- AHMADI, S., BLUM, A., AND YANG, K. 2023. Fundamental bounds on online strategic classification. In *Proceedings of the 24th ACM Conference on Economics and Computation*. 22–58.
- AHMADI, S., YANG, K., AND ZHANG, H. 2024. Strategic littlestone dimension: Improved bounds on online strategic classification. *arXiv preprint arXiv:2407.11619*.
- ALON, T., DOBSON, M., PROCACCIA, A., TALGAM-COHEN, I., AND TUCKER-FOLTZ, J. 2020. Multiagent evaluation mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 1774–1781.
- BEHAVOD, Y., LIGETT, K., WU, S., AND ZIANI, J. 2021. Gaming helps! learning from strategic interactions in natural dynamics. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1234–1242.
- BEHAVOD, Y., PODIMATA, C., WU, S., AND ZIANI, J. 2022. Information discrepancy in strategic learning. In *International Conference on Machine Learning*. PMLR, 1691–1715.
- BJÖRKEGREN, D., BLUMENSTOCK, J. E., AND KNIGHT, S. 2020. Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*.
- BRAVERMAN, M. AND GARG, S. 2020. The role of randomness and noise in strategic classification. In *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference)*, A. Roth, Ed. LIPIcs, vol. 156. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 9:1–9:20.
- CHANG, T., WARRENBURG, L., PARK, S.-H., PARIKH, R. B., MAKAR, M., AND WIENS, J. 2024. Who’s gaming the system? a causally-motivated approach for detecting strategic adaptation. *arXiv preprint arXiv:2412.02000*.

¹¹Another promising application domain is the health insurance industry, as recently discussed in [Chang et al. 2024].

- CHEN, Y., LIU, Y., AND PODIMATA, C. 2020. Learning strategy-aware linear classifiers. *Advances in Neural Information Processing Systems* 33, 15265–15276.
- CHEN, Y., PODIMATA, C., PROCACCIA, A. D., AND SHAH, N. 2018. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 9–26.
- COHEN, L., MANSOUR, Y., MORAN, S., AND SHAO, H. 2024. Learnability gaps of strategic classification. *arXiv preprint arXiv:2402.19303*.
- COHEN, L., SHARIFI-MALVAJERDI, S., STANGL, K., VAKILIAN, A., AND ZIANI, J. 2024. Bayesian strategic classification. *arXiv preprint arXiv:2402.08758*.
- DEE, T. S., DOBBIE, W., JACOB, B. A., AND ROCKOFF, J. 2019. The causes and consequences of test score manipulation: Evidence from the new york regents examinations. *American Economic Journal: Applied Economics* 11, 3, 382–423.
- DEKEL, O., FISCHER, F., AND PROCACCIA, A. D. 2010. Incentive compatible regression learning. *Journal of Computer and System Sciences* 76, 8, 759–777.
- DONG, J., ROTH, A., SCHUTZMAN, Z., WAGGONER, B., AND WU, Z. S. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 55–70.
- DRANOVE, D., KESSLER, D., MCCLELLAN, M., AND SATTERTHWAITE, M. 2003. Is more information better? the effects of “report cards” on health care providers. *Journal of political Economy* 111, 3, 555–588.
- EBRAHIMI, R., VACCARO, K., AND NAGHIZADEH, P. 2024. The double-edged sword of behavioral responses in strategic classification: Theory and user studies. *arXiv preprint arXiv:2410.18066*.
- ESTORNELL, A., DAS, S., LIU, Y., AND VOROBAYCHIK, Y. 2023. Group-fair classification with strategic agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 389–399.
- GHALME, G., NAIR, V., EILAT, I., TALGAM-COHEN, I., AND ROSENFELD, N. 2021. Strategic classification in the dark. In *International Conference on Machine Learning*. PMLR, 3672–3681.
- GONZALEZ-LIRA, A. AND MOBARAK, A. M. 2019. Slippery fish: Enforcing regulation under subversive adaptation. Tech. rep., IZA Discussion Papers.
- GREENSTONE, M., HE, G., JIA, R., AND LIU, T. 2022. Can technology solve the principal-agent problem? evidence from china’s war on air pollution. *American Economic Review: Insights* 4, 1, 54–70.
- GUPTA, V., NOKHIZ, P., ROY, C. D., AND VENKATASUBRAMANIAN, S. 2019. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*.
- HAGHTALAB, N., IMMORLICA, N., LUCIER, B., AND WANG, J. Z. 2020. Maximizing welfare with incentive-aware evaluation mechanisms. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, Ed. ijcai.org, 160–166.
- HARDT, M., MEGIDDO, N., PAPADIMITRIOU, C., AND WOOTTERS, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*. 111–122.
- HARRIS, K., AGARWAL, A., PODIMATA, C., AND WU, Z. S. 2024. Strategyproof decision-making in panel data settings and beyond. *ACM SIGMETRICS Performance Evaluation Review* 52, 1, 69–70.
- HARRIS, K., HEIDARI, H., AND WU, S. Z. 2021. Stateful strategic regression. *Advances in Neural Information Processing Systems* 34, 28728–28741.
- HARRIS, K., NGO, D. D. T., STAPLETON, L., HEIDARI, H., AND WU, S. 2022. Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In *International Conference on Machine Learning*. PMLR, 8502–8522.
- HARRIS, K., PODIMATA, C., AND WU, S. Z. 2023. Strategic apple tasting. *Advances in Neural Information Processing Systems* 36, 79918–79945.
- HAUPT, A., HADFIELD-MENELL, D., AND PODIMATA, C. 2023. Recommending to strategic users. *arXiv preprint arXiv:2302.06559*.

- HOROWITZ, G. AND ROSENFELD, N. 2023. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*. PMLR, 13233–13253.
- HOROWITZ, G., SOMMER, Y., KOREN, M., AND ROSENFELD, N. 2024. Classification under strategic self-selection. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- HU, L., IMMORLICA, N., AND VAUGHAN, J. W. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 259–268.
- JAGADEESAN, M., MENDLER-DÜNNER, C., AND HARDT, M. 2021. Alternative microfoundations for strategic classification. In *International Conference on Machine Learning*. PMLR, 4687–4697.
- KARIMI, A.-H., BARTHE, G., SCHÖLKOPF, B., AND VALERA, I. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.
- KLEINBERG, J. AND RAGHAVAN, M. 2020. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)* 8, 4, 1–23.
- KRISHNASWAMY, A. K., LI, H., REIN, D., ZHANG, H., AND CONITZER, V. 2021. Classification with strategically withheld data. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5514–5522.
- LECHNER, T. AND URNER, R. 2022. Learning losses for strategic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 7337–7344.
- LECHNER, T., URNER, R., AND BEN-DAVID, S. 2023. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*. PMLR, 18714–18732.
- LEVANON, S. AND ROSENFELD, N. 2021. Strategic classification made practical. In *International Conference on Machine Learning*. PMLR, 6243–6253.
- MILLER, J., MILLI, S., AND HARDT, M. 2020. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*. PMLR, 6917–6926.
- MILLI, S., MILLER, J., DRAGAN, A. D., AND HARDT, M. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 230–239.
- PEARL, J. 2009. *Causality*. Cambridge university press.
- PERDOMO, J., ZRNIC, T., MENDLER-DÜNNER, C., AND HARDT, M. 2020. Performative prediction. In *International Conference on Machine Learning*. PMLR, 7599–7609.
- ROSENFELD, E. AND ROSENFELD, N. 2024. One-shot strategic classification under unknown costs. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- SHAO, H., BLUM, A., AND MONTASSER, O. 2024. Strategic classification under unknown personalized manipulation. *Advances in Neural Information Processing Systems* 36.
- SHAVIT, Y., EDELMAN, B., AND AXELROD, B. 2020. Causal strategic linear regression. In *International Conference on Machine Learning*. PMLR, 8676–8686.
- SPIELMAN, D. A. AND TENG, S.-H. 2009. Smoothed analysis: an attempt to explain the behavior of algorithms in practice. *Communications of the ACM* 52, 10, 76–84.
- SUNDARAM, R., VULLIKANTI, A., XU, H., AND YAO, F. 2023. Pac-learning for strategic classification. *Journal of Machine Learning Research* 24, 192, 1–38.
- TSIRTSIS, S. AND GOMEZ RODRIGUEZ, M. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems* 33, 16749–16760.
- VALIANT, L. G. 1984. A theory of the learnable. *Communications of the ACM* 27, 11, 1134–1142.
- ZHANG, H. AND CONITZER, V. 2021. Incentive-aware pac learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 5797–5804.
- ZRNIC, T., MAZUMDAR, E., SASTRY, S., AND JORDAN, M. 2021. Who leads and who follows in strategic classification? *Advances in Neural Information Processing Systems* 34, 15257–15269.