Strategic Adaptation to Classifiers: A Causal Perspective

John Miller Smitha Milli Moritz Hardt

November 4, 2019

Abstract

Consequential decision-making incentivizes individuals to adapt their behavior to the specifics of the decision rule. A long line of work has therefore sought to understand and anticipate adaptation, both to prevent strategic individuals from "gaming" the decision rule and to explicitly motivate individuals to improve. In this work, we frame the problem of adaptation as performing interventions in a causal graph. With this causal perspective, we make several contributions. First, we articulate a formal distinction between gaming and improvement. Second, we formalize strategic classification in a new way that recognizes that the individual may improve, rather than only game. In this setting, we show that it is beneficial for the decision-maker to incentivize improvement. Third, we give a reduction from causal inference to designing incentivizes for improvement. This shows that designing good incentives, while desirable, is at least as hard as causal inference.

1 Introduction

When classifiers are used for consequential decision-making, individuals will inevitably seek to improve their classification outcome. News sites optimize their headlines to increase their search rankings. Borrowers open more credit lines to increase their chances of getting a low interest rate loan. In consequential classification, *adaptation* is universal. The way adaptation is typically modelled, however, is not.

A long line of work has framed adaptation as a form of *gaming*, that is, undesirable strategic behavior that can diminish utility for the decision maker (Hardt et al., 2016; Dong et al., 2018; Dalvi et al., 2004; Brückner et al., 2012). Gaming and its consequences are at the heart of Goodhart's law, and failure to mitigate gaming can have serious consequences in a range of applications, from abuse detection on online platforms to credit scoring in financial applications (Strathern, 1997). Work in *strategic classification* has therefore sought to prevent adaptation by virtue of more conservative decision boundaries (Brückner et al., 2012; Hardt et al., 2016; Dong et al., 2018).

At the same time, recent work has rightfully recognized that adaptive behavior can also correspond to actual individual *improvement* (Bambauer and Zarsky, 2018; Kleinberg and Raghavan, 2019). Rather than tweak their headlines to improve search rankings, new sites can also boost their rankings by creating better original content. We should not aim to prevent an individual from increasing their net worth in an effort to get a loan, or solving math exercises to improve their SAT performance. Kleinberg and Raghavan (2019) consequently focuses on designing classifiers that *incentivize* individual improvement. Similarly, work on both counterfactual explanations (Wachter et al., 2017) and actionable recourse (Ustun et al., 2019) is

motivated by helping individuals improve their classification outcomes and implicitly conceptualizes adaptation as improvement.

Both of these lines of work assume at the outset that adaptation is *either* gaming *or* improvement. In practice, both gaming and improvement can occur, and mistaking one for the other can lead to perverse results. If the decision-maker assumes all adaptation is gaming, strategic classification is a natural solution concept. However, where strategic classification methods may be robust to adaptation, they also make it more difficult for individuals to improve. At the same time, if the decision-maker assumes that all adaptation is improvement, ill-considered explanations and unanticipated incentives can unintentionally encourage gaming. Imagine if an automated grading system provided counterfactual explanations like: "You can improve the score of your essay by including more words that have more than seven characters."

If adaptation is sometimes gaming and sometimes improvement, how do we formally distinguish between the two? Moreover, from the perspective of the decision maker, how do we incentivize improvement without encouraging gaming? This leads to the problem of *incentive design*. Although a topic with much recent interest, the fundamental difficulties involved with designing classifiers that incentivize improvement are not yet well-understood.

A causal framework. In this work, we develop a *causal* perspective on strategic adaptation that distinguishes between gaming and improvement and sheds light on the fundamental difficulty of designing classifiers that incentivize improvement. The causal framework allows us to reason about individual adaptation without making a priori assumptions on whether adaptation is gaming or improvement.

We conceptualize individual adaptation as performing an *intervention* in a causal model that includes all relevant features X, a predictor $\hat{Y}$, as well as the target variable Y. Some features are causal for Y in that they are ancestors of Y in the causal graph. The rest are *non-causal* features that are merely correlated with Y.

We use the causal structure to formally distinguish between improvement and gaming. For intuition on the role of causality, consider the difference between non-causal and causal features. When individuals adapt non-causal features they cannot change the target variable Y, but they might nonetheless influence the predictor $\hat{Y}$. Hence, adaptation of non-causal features can only be gaming but not improvement. On the other hand, adaptation of causal features can lead to improvement because changes to causal features can influence both the target variable and the predictor at the same time.

The distinction between causal and non-causal adaptation in a classification setting is intuitive. In fact, such considerations were present, for example, in early work on statistical risk assessment in lending (Hand et al., 1997). However, to understand what a classifier *incentivizes*—improvement or gaming—we must also understand how the causal model interacts with the *agent model*, the model of how individuals adapt. We discuss such subtleties in Section 3.

Applying the framework. Traditionally, work in strategic classification assumes all adaptation is gaming because it assumes individual adaptation cannot change the target variable Y. Using our causal framework, we introduce a *dynamic* variant of the strategic classification problem in which the target variable changes in response to individual adaptation. This formalism simultaneously captures both gaming and improvement behaviors. Reasoning about the dynamic classification problem, we prove adaptation need not be at odds with the ob-

jectives of purely profit-maximizing institutions. Indeed, profit-maximizing institutions can increase their utility by explicitly *incentivizing improvement*.

We then employ our causal framework to study the hardness of designing classifiers that incentivize improvement. Informally, we prove that designing classifiers with *good incentives* is as hard as causal inference. Formally, we prove that any oracle that decides whether a classifier exists that incentivizes improvement can be leveraged to orient the edges of a causal graph. In other words, causal reasoning is unavoidable when addressing problems of incentive design.

At a high level, our work demonstrates causality is crucial for understanding strategic adaptation. Teasing apart the difference between gaming and improvement relies on a causal framework, and our reduction shows that causal analysis is inevitably required to develop classifiers that incentivize improvement rather than gaming. Consequently, properly reasoning about strategic adaptation thus necessitates reasoning about causality.

2 Causal background

We use the language of *structural causal models* (Pearl, 2009) as a formal framework for causality. A structural causal model (SCM) consists of endogenous variables $X = (X_1, \ldots, X_n)$, exogenous variables $U = (U_1, \ldots, U_n)$, a distribution over the exogenous variables, and a set of structural equations that determine the values of the endogenous variables. The structural equations can be written

$$X_i = g_i(\mathbf{PA}_i, U_i), \quad i = 1, \ldots, n,$$

where g_i is an arbitrary function, $\mathbf{PA}_i$ represents the other endogenous variables that determine X_i, and U_i represents exogenous noise due to unmodeled factors.

Every structural causal model gives rise to a *causal graph* where a directed edge exists from X_i to X_j if X_i is an input to the structural equation governing X_j, i.e. $X_i \in \mathbf{PA}_j$. If the causal graph is a directed acyclic graph (DAG), we call it a *causal DAG*. We restrict ourselves to considering only *Markovian* structural causal models, models which have an acyclic causal graph and exogenous variables which are independent from one another.

An *intervention* is a modification to the structural equations of an SCM. For example, an intervention may consist of replacing the structural equation $X_i = g_i(\mathbf{PA}_i, U_i)$ with a new structural equation $X_i := x_i$ that holds X_i at a fixed value. We use $:=$ to denote modifications of the original structural equations. When the structural equation for one variable is changed, other variables can also change. Suppose Z and X are two endogenous nodes, Then, we use the notation $Z_{X:=x}$ to refer to the variable Z in the modified SCM with structural equation $X := x$.

Given the values u of the exogenous variables U, the endogenous variables are completely deterministic. We use the notation $Z(u)$ to represent the deterministic value of the endogenous variable when the exogenous variables U are equal to u. Similarly, $Z_{X:=x}(u)$ is the value of Z in the modified SCM with structural equation $X := x$ when $U = u$.

More generally, given some event E, $Z_{X:=x}(E)$ is the random variable Z in the modified SCM with structural equations $X := x$ where the distribution of exogenous variables U is updated by conditioning on the event E. For more details on this *counterfactual* notion, see (Pearl, 2009).

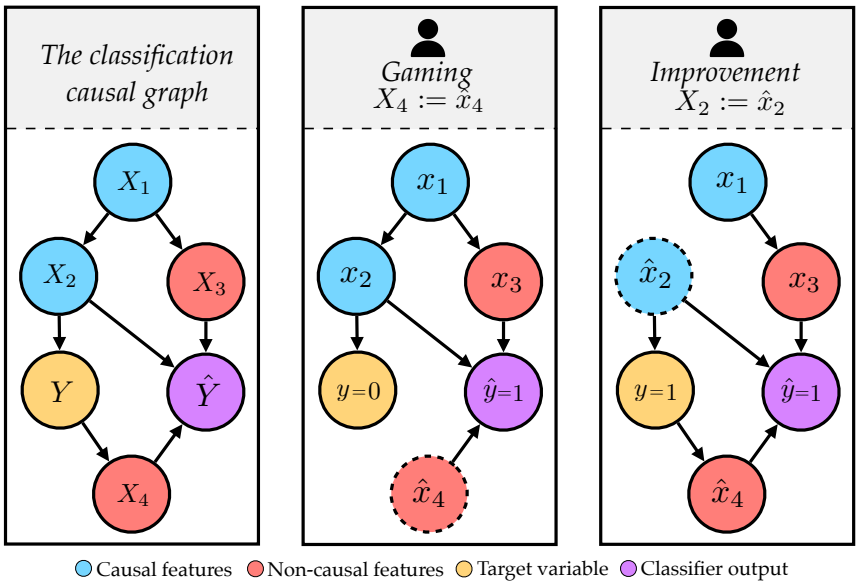

Figure 1: Illustration of the causal framework for reasoning about strategic adaptation. Strategic adaptation corresponds to interventions in the causal graph. Gaming is interventions that change the classification $\hat{y}$, but do not change the true label y. Improvement, on the other hand, consists of interventions that change both the classification $\hat{y}$ and the true label y.

3 A Causal Framework for Strategic Adaptation

When machine learning classifiers are used to make decisions that people care about, people will strategically *adapt* their features in order to get their desired classification. In this section, we put forth a causal framework for modeling strategic adaptation. We argue that the causal perspective is what is needed to distinguish between two prior ways of viewing adaptation — as "gaming" or as "improvement".

3.1 Gaming or improvement?

In machine learning, strategic adaptation has historically been viewed in a negative, adversarial light. A motivating example is often a spammer who makes small modifications to their email in order to get past a spam filter. Or a content creator who tweaks their website in order to rank higher in search engine results. Such cases are sometimes referred to as "gaming" the classifier. Inspired by cases of gaming, when it comes to strategic adaptation, research has mainly focused on creating more "robust" classifiers, ones in which it is harder for an individual to change their classification outcome through adapting their features (Hardt et al., 2016; Dong et al., 2018).

Gaming is not, however, all there is to adaptation. Imagine that a software company begins using an algorithm to filter or source software engineering job applicants (such algorithms are indeed used by many large tech companies). Compared to traditional interview processes, the algorithm they use weighs the amount of open source contributions a candidate has more heavily. Some individuals realize this and *adapt*—they begin focusing more of their energy on making open source contributions. Is this gaming? If the contributions these individuals made

were only stylistic and created for the sole purpose of increasing their contribution count, then yes. But, if the contributions they made required them to actually engage with the open-source projects, then it may have been a pedagogically valuable experience. The extra effort they put into making open-source contributions may have actually made them a better programmer, and thus a better candidate.

The gaming view of adaptation overlooks the fact that it is possible for individuals to *improve* through adaptation. The spammer never modifies their desired email into a non-spam email because they are committed to sending spam. In contrast, an individual seeking a job is not opposed to improving their skill level in order to do so. In practice, both gaming and improvement can happen, and any complete approach to addressing the incentives produced by machine learning systems must be able to distinguish between them.

3.2 A causal distinction

The key to distinguishing between gaming and improvement is understanding the structural causal model (Section 2) that underlies the classification problem.

Recall that a structural causal model has two types of nodes: endogenous nodes and exogenous nodes. The endogenous nodes are the individual's true label Y, their features $X = \{X_1, \dots, X_n\}$, and their classification outcome $\hat{Y}$. The structural equation for $\hat{Y}$ is represented by the classifier $\hat{Y} = f(Z)$, where $Z \subseteq X$ are the features that the classifier f has access to and uses. The exogenous variables U represent all of the other unmodeled factors.

Individuals are not static — they may adapt some of their features in order to change their classification outcome. The agent model $\Delta(x, f)$ defines how individuals adapts. The agent model takes in an individual's features $x = (x_1, \dots, x_n)$, as well as the classifier f, and returns a new set of features $(x'_1, \dots, x'_n)$. Note that, as a function of u, $\Delta(x; f)$ is deterministic.

Critically, the individual's adaptation can change their label. To formally define the label after adaptation, let $A = \{i : \Delta(x, f)_i \neq x_i\}$ be the subset of features the individual adapts, and let X_A index those features. In the context $U = u$, the value of the true label after adaptation is given by $Y_{X_A := \Delta(x,f)_A}(u)$. The dependence on A ensures that, if an individual only intervenes on a subset of features, the remaining features are still consistent with the original causal model. For brevity, however, we frequently omit reference to A and write $Y_{X := \Delta(x,f)}(u)$.

In general, the individual only observes and adapts based on features X, and the exogenous variables U many contain other factors, for instance the state of the economy, which are not properly thought of as part of the individual. To define improvement and gaming, we therefore instead consider the labels $Y_{X := \Delta(x,f)}(\{X = x\})$, which retain randomness over background factors not determined by the observed features $X = x$.

With these quantities in hand, we can now formally define improvement and gaming.

Definition 3.1. We say that an individual with features $X = x$ *improves* if their adaptation leads to improving their true label on average:

$$\mathbb{E}\Big[Y_{X := \Delta(x,f)}(\{X = x\})\Big] - \mathbb{E}[Y] > 0.$$

Otherwise, we say that the individual *games*.

Extending this definition to the population level, a classifier *incentivizes improvement* if individuals improve on average.

Definition 3.2. We say that a classifier *incentivizes improvement* if individuals (in expectation over the population) improve:

$$\mathbb{E}_X\left[\mathbb{E}\left[Y_{X:=\Delta(x,f)}(\{X=x\})\right]\right]-\mathbb{E}[Y]>0$$

Otherwise, if a classifier does not incentivize improvement, we say that a classifier *incentivizes gaming*.

While the distinction between $Y_{X:=\Delta(x,f)}(\{X=x\})$ and $Y_{X:=\Delta(x,f)}(u)$ matters greatly at the individual level, it is less important at the population level. Indeed, by the tower property, Definition (3.2) is equivalent to asking whether a classifier incentivizes improvement on average across contexts U. In particular, a classifier f incentivizes improvement if and only if $\mathbb{E}_U\left[Y_{X:=\Delta(x,f)}(U)\right]-\mathbb{E}_U[Y]>0$. In the sequel, we make use of both representations.

In Definition (3.2), whether a classifier incentivizes improvement or gaming depends on both (a) the causal model and (b) the agent model $\Delta(x,f)$. We next elaborate on both of these models in turn.

3.3 The causal model

To understand the role of the causal model, let us revisit our example of the company seeking to hire software engineers. The company has observed that software engineering skill Y is positively correlated to experience contributing to open source projects X. There are two possible causal DAGs to explain this correlation. In one scenario, we have that $Y \to X$: the more skilled one becomes, the more likely one is to contribute to open-source projects, perhaps because contributing to open-source requires a baseline level of knowledge. In the other scenario, $X \to Y$: the more someone contributes to open source, the more skilled they become.

Now, the company begins using an algorithm that positively weighs contributions to open source. Some candidates realize this and adapt; they begin contributing more to open source. Which causal DAG is true is crucial for distinguishing between whether the candidates are gaming or are improving. In the first world, open source contributions reflect, but do not cause, a candidate's skill level. Strategic individuals in this world are merely gaming the classifier. In the second world, contributing to open source causes a candidate's skill level to increase. Strategic individuals in this world are actually improving their skill.

Thus, determining whether an intervention is gaming or improvement requires knowledge of the causal relationships in the problem. When the individual intervenes on non-causal features, there is no way for them to improve their true label Y. Thus, changes to non-causal features are mere gaming. On the other hand, changes to causal features can affect the true label Y, potentially leading to improvement.

3.4 The agent model

The causal model is essential because it determines which interventions lead to improvement or gaming. At the same time, the agent model is essential because it determines which interventions the agent executes in the first place.

To understand the role of the agent model, let us consider another example – a toy version of search engine optimization (SEO). A search engine is trying to predict the quality of a website Y. The feature X_1, the uniqueness of the website's content, is causal for Y. The search engine cannot directly observe X_1, however they can observe the non-causal feature X_2, the number

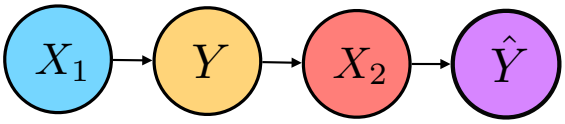

Figure 2: Reasoning about the agent model is essential for determining whether a classifier incentivizes improvement or gaming. Even though the classification $\hat{Y}$ only depends on X_2, the agent can change the label by manipulating either X_1 or X_2. The causal model determines whether the agent's change is gaming or improvement, but the agent model determines which actions the agent actually takes.

of incoming links to the website. They design a classifier $f(X_2)$ that outputs a prediction $\hat{Y}$ of a website's quality. The causal model is illustrated in Figure 2.

There are two ways that websites may improve their prediction $\hat{Y} = f(X_2)$. One way is for the website to *improve* by creating more unique content. Creating more unique content improves the quality of the website Y, which will also increases the number of incoming links, and thus also improve the predictor $\hat{Y}$. The other way is to *game* and directly increase the links, by say, buying them. Which type of adaptation occurs in response to the classifier f depends on the agent model $\Delta(x, f)$.

Although our formal framework accommodates a variety of agent models, in this work, as in prior work on strategic classification, we assume the individual *best responds* to the classifier (Hardt et al., 2016; Dong et al., 2018). Following Ustun et al. (2019), in response to the classifier f, an individual with features x takes an action a to change her features to $x+a$. However, this modification comes at a cost, $c(a; x)$, and the action the individual takes is determined by balancing the benefits of classification $f(x+a)$ with the cost of adaptation $c(a; x)$.

Definition 3.3 (Best response agent model)**.** Given a cost function $c : \mathcal{A} \times \mathcal{X} \to \mathbf{R}_+$ and a classifier $f : \mathcal{X} \to \{0, 1\}$, an individual with features x best responds to the classifier by moving to features $\Delta(x, f)$, where

$$a^* = \arg\max_a f(x+a) - c(a; x) \quad \text{and} \quad \Delta(x, f) = x + a^*.$$

When clear from context, we omit the dependence on f and write $\Delta(x)$.

A key implication of the best response agent model is that, when there are multiple paths to improving one's classification, the individual will take the lowest cost path. Returning to the SEO example, we imagine that in many cases it would be easier for corporations to increase the number of incoming links to their website than to create more unique content. Indeed, JCPenny once boosted their Google rankings by buying thousands of links from unrelated sites. Google had to temporarily apply 'manual action' to reduce JCPenny's rankings (Ziewitz, 2019).

We can also imagine scenarios that have the same causal graph as the SEO example (Figure 2), but in which people are very unlikely to game. Imagine an online marketplace for freelancers is attempting to predict Y the quality of a freelancer (see e.g. Kokkodis et al. (2015)). Let X_1 denote how much time a freelancer spends on a project, which is causal to Y. However, X_1 is not directly observable by the company. The company can instead measure the non-causal feature like ratings of the freelancer that are given by verified employers, which we denote X_2. Since the ratings X_2 can only be given by verified employers, it is impossible (or at least very high cost) for a freelancer to manipulate X_2 directly. Thus, although the causal structure is

exactly the same as the SEO example, we would expect freelancers to improve rather than to game. More generally, if the non-causal feature X_2 and the target label Y have a common cause, then including feature X_2 in the model can still incentivize improvement if it is more difficult to change X_2 than to change upstream causal features that affect both Y and X_2.

Together, the causal graph and agent model provide a heuristic guideline for designing classifiers that incentivize improvement: prioritize causal features over non-causal features. Improvement only happens when an intervention changes causal features, so a classifier that emphasizes causal features more likely to lead to improvement. As we have seen from the freelancer example, technically, even a classifier that uses non-causal features, can produce incentives for improvement. However, whether or not the classifier incentivizes improvement depends on whether it is lower-cost for the agent to change non-causal features or causal features. In most scenarios, we would posit that it is easier to change non-causal features than causal ones. Thus, the use of non-causal features for improvement should spark caution.

4 Adaptation need not be at odds with profit-maximization

In this section, we consider strategic adaptation from the perspective of the profit-maximizing institution. In machine learning, adaptation is typically modeled purely as gaming and thus harmful to institutional objectives, regardless of the decision-rule or the types of features it uses (Dalvi et al., 2004; Hardt et al., 2016). Consequently, institutions often attempt to limit adaptation by keeping the decision-rule secret or picking models that are ostensibly robust to adaptation (Tramèr et al., 2016; Hardt et al., 2016; Dong et al., 2018).

As the previous section suggests, however, adaptation is not necessarily at odds with the institution's objectives. In many important cases, when people improve rather than game the classifier, strategic adaptation also benefits the institution. For instance, when someone adapts to a fitness tracker (O'Neill, 2018) by exercising more, this improves her health *and* lowers the insurance company's risk. Similarly, if a loan applicant reduces her outstanding debt to obtain a loan, the institution benefits from her corresponding higher probability of repayment.

In these cases, drawing an actionable distinction between gaming and improvement shifts the institution's emphasis from one of preventing adaptation to one of *incentivizing improvement*. Rather than keep the model secret, the institution might focus on what parts of the model to reveal so that adaptation is improvement rather than gaming (Homonoff et al., 2019). Similarly, rather than design classifiers that are robust to adaptation, the institution might search for decision-rules that encourage improvement.

In the rest of this section, we introduce a measure of *dynamic utility* that explicitly accounts for improvement. Then, we prove strategic adaptation is not necessarily at odds with institutional objectives. Indeed, under dynamic utility, the optimal strategy for the institution is to use decision-rules that explicitly incentivize improvement.

4.1 Dynamic utility

Prior work in strategic classification uses models of institutional utility assume that an individual's adaptation is always gaming. In this section, we introduce a measure of *dynamic utility* that captures the fact that individuals may actually improve through their adaptation.

Without adaptation, the institution maximizes the static utility

$$\mathcal{U}_{\text{static}}(f) = \mathbb{P}\{f(X) = Y\}.$$

In the strategic setting, negatively classified individuals then attempt to change their classification by best-responding and changing their features from x to some $\Delta(x; f)$. Typically, in strategic classification (Hardt et al., 2016), the institution maximizes

$$\mathcal{U}_{\text{strategic}}(f) = \mathbb{P}\{f(\Delta(X)) = Y\}.$$

However, this model does not take into account how changing X to $\Delta(X)$ can also affect Y. To remedy this, we consider a *dynamic* measure of utility

$$\mathcal{U}_{\text{dynamic}}(f) = \mathbb{E}_X\Big[\mathbb{P}\Big\{f(\Delta(x)) = Y_{\mathbf{X}:=\Delta(x)}(\{X = x\})\Big\}\Big].$$

Dynamic utility explicitly allows for individual adaptation. For instance, suppose a previously uncreditworthy loan applicant becomes eligible for a loan by increasing her net worth, which in turn increases her probability of repayment. Both her classification outcome $f(\Delta(x))$ and label $Y_{\mathbf{X}:=\Delta(x)}(\{X = x\})$ change. While $\mathcal{U}_{\text{dynamic}}$ accounts for this adaptation, $\mathcal{U}_{\text{strategic}}$ actually penalizes the classifier for the individual's improvement!

4.2 Dynamic utility incentivizes improvement

We now turn to showing that when institutions optimize dynamic utility, they are naturally incentivized to pursue policies that lead to improvement. As a running example, consider a lending scenario where the institution predicts whether an individual will repay a loan, $Y \in \{0, 1\}$, on the basis of features $x \in \mathcal{X}$ using some classifier $f : \mathcal{X} \to \{0, 1\}$. Suppose the institution earns 1 unit of profit for true positives and suffers κ units of loss for false positives.

Since dynamic utility is a population concept, rather than condition on $\{X = x\}$, by the tower propery, we can equivalently take expectations directly over the exogenous variables, which yields

$$\mathcal{U}_{\text{dynamic}}(f) = \mathbb{P}_U\Big\{f(\Delta(X)) = 1, Y_{\mathbf{X}:=\Delta(x)}(U) = 1\Big\} - \kappa\mathbb{P}_U\Big\{f(\Delta(X)) = 1, Y_{\mathbf{X}:=\Delta(x)}(U) = 0\Big\}.$$

We state our main result in terms of the *rate of improvement* of individuals who initially receive a negative classification. In particular, for $y \in \{0, 1\}$ define

$$I_y = \mathbb{P}_U\Big\{Y_{\mathbf{X}:=\Delta(x)}(U) = 1 \mid Y(U) = y, f(X) = 0, f(\Delta(X)) = 1\Big\}.$$

Thus, I_0 corresponds to the rate of improvement for true negatives, and I_1 corresponds to rate of improvement (or lack of change) for false negatives. Informally, we prove higher rates of I_0 and I_1 lead to higher dynamic utility, and, when the rate of improvement is sufficiently high, the dynamic utility can actually be better than static utility. Ceteris paribus, institutions are incentivized to use classifiers with higher rates of improvement.

Proposition 4.1. *Assume the base rate $\mathbb{P}\{Y = 1\} \in (0, 1)$ and some fraction of the true negatives can change their label. For any trade-off cost κ,*

1. *$\mathcal{U}_{\text{dynamic}}$ is monotonically increasing in I_0 and I_1.*
2. *if $I_0 > \frac{\kappa}{1+\kappa}$ and $I_1 > \frac{\kappa}{1+\kappa}$, then $\mathcal{U}_{\text{dynamic}}(f) > \mathcal{U}_{\text{static}}(f)$.*

Proof. For a fixed classifier f, let $p_{ab} = \mathbb{P}\{Y = a, f(X) = b\}$ for $a, b \in \{0, 1\}$. In terms of these quantities, the static utility is given by

$$\mathcal{U}_{\text{static}}(f) = p_{11} - \kappa p_{01}.$$

To compute dynamic utility, we consider the individuals who strategically adapt. In the best-response model, only individuals initially with $f(x) = 0$ will adapt their features. Thus, there are two groups to consider:

1. True negatives ($f(x) = 0, y = 0$): Let γ_0 denote fraction of this group manages to change their classification, i.e.

$$\gamma_0 = \mathbb{P}\{f(\Delta(X)) = 1 \mid f(X) = 0, Y = 0\}.$$

2. False negatives ($f(x) = 0, y = 1$): Similarly, let γ_1 denote the fraction of this group changes their classification.

By assumption, γ_0 is strictly positive. In terms of the aforementioned quantities, the dynamic utility is given by

$$\mathcal{U}_{\text{dynamic}}(f) = (p_{11} + p_{10}\gamma_1\left[(1+\kappa)I_1 - \kappa\right] + p_{00}\gamma_0\left[(1+\kappa)I_0 - \kappa\right] - \kappa p_{01}).$$

Simple inspection shows $\mathcal{U}_{\text{dynamic}}$ is monotonically increasing in I_0, I_1. To conclude, if $I_0 > \frac{\kappa}{1+\kappa}$ and if $I_1 > \frac{\kappa}{1+\kappa}$, then

$$\mathcal{U}_{\text{dynamic}}(f) \quad = p_{11} + p_{10}\gamma_1\left[(1+\kappa)I_1 - \kappa\right] + p_{00}\gamma_0\left[(1+\kappa)I_0 - \kappa\right] - \kappa p_{01} > p_{11} - \kappa p_{01} = \mathcal{U}_{\text{static}}(f).$$

□

The key phenomenon underlying Proposition (4.1) is the institution benefits when many individuals adapt and change their classification *and* this adaptation is improvement rather than gaming.

To make this point clear, suppose the institution can perfectly classify individuals. In the static case without adaptation, the optimum $\mathcal{U}_{\text{static}}$ is simply the base rate of positive individuals $\mathbb{P}\{Y = 1\}$. In the strategic setting, $\mathcal{U}_{\text{strategic}}$ may in general be less than the base rate if true negatives game the classifier. However, under dynamic utility, the optimal policy for the institution is to deploy a classifier so that (1) all of the true negatives ($Y(u) = 0$) are correctly classified ($f(X) = 0$), and (2) they all successfully adapt ($f(\Delta(X)) = 1$) *and* improve their label ($Y_{X:=\Delta(x)}(u) = 1$). In this case, the optimum value $\mathcal{U}_{\text{dynamic}}$ is 1, which is greater than $\mathcal{U}_{\text{static}}$. Furthermore, in this setting, any classifier that maximizes dynamic utility must necessarily maximize the total improvement incentivized by the classifier, as the following corollary makes precise.

Corollary 4.1. *Suppose there exists classifiers with $\mathbb{P}\{f(X) = Y\} = 1$. Suppose further every setting (I_0, I_1) is achievable. If f^* maximizes dynamic utility, i.e. $f^* \in \arg\max_f \mathcal{U}_{\text{dynamic}}(f)$, then f^* also maximizes improvement,*

$$f^* \in \arg\max_f \; \mathbb{E}_X \mathbb{E}\Big[Y_{X:=\Delta(x)}(\{X = x\})\Big] - \mathbb{E}[Y].$$

5 Giving good incentives is as hard as causal inference

Many types of decision-makers may desire to incentivize improvement. For instance, Kleinberg and Raghavan (2019) examine the scenario where an evaluator, e.g. a teacher, wants to construct a grading system that incentivizes students to study and learn the course material rather than cheat and look up answers online. In the previous section, we showed that in the dynamic setting, a profit-maximizing institution can also benefit from incentivizing improvement. So, we may now wonder, how hard is it to create classifiers that incentivize improvement? In this section, we prove a reduction from causal inference to designing good incentives, showing that creating classifiers that incentivize improvement is as at least as hard as causal inference.

As illustrated in Section 3, in order to evaluate the incentives produced by a classifier, we need knowledge of two things: the causal model and the user model. We need to know (1) which actions lead to improvement and (2) whether or not users can be incentivized to take those actions. Kleinberg and Raghavan (2019) focus on the second problem; they determine whether or not a *given* set of actions is incentivizable. However, this masks the difficulty of the step that comes before that – figuring out which actions are good in the first place.

Since adaptation of non-causal features can never lead to improvement, figuring out which actions lead to improvement or not necessitates distinguishing between causal and non-causal features. In other words, any procedure that can provide incentives for improvement must capture some, possibly implicit, knowledge about the causal relationship between the features and the label.

The main result of this section generalizes this intuition and, in doing so, reveals the true difficulty of incentive design. We establish a reduction from orienting the edges in a causal graph to designing classifiers that incentivize improvement. Orienting the edges in a causal graph is not generally possible from observational data alone (Peters et al., 2017), though it can be addressed through active intervention (Eberhardt et al., 2005). Therefore, any procedure for constructing classifiers that incentivize improvement must at its core also solve a non-trivial causal inference problem.

5.1 The good incentives problem

We first formally state the problem of designing classifiers with good incentives. As a running example, consider the hiring example presented in Section 3. A decision-maker has access to a distribution over features (open-source contributions, employment history, coding test scores, etc), a label (engineering ability), and wishes to design a decision rule that incentivizes strategic individuals to improve their engineering ability rather than game the classifier. As discussed in Section 3, the decision-maker must reason about the agent model governing adaptation, and we assume the individual *best responds* to the classifier according to some cost. All together, this leads to the *good incentives problem*.

Definition 5.1 (Good Incentives Problem)**.** Given a cost function $c : \mathcal{A} \times \mathcal{X} \to \mathbf{R}_+$, and a joint distribution $P_{X,Y}$ over a set of features X and a label Y, and assuming that individuals best respond to a classifier f so that

$$a^* = \arg\max_a f(x+a) - c(a;x) \quad \text{and} \quad \Delta(x,f) = x + a^*.$$

The Good Incentives problem is to determine whether or not there exists a classifier f that incentivizes improvement: $\mathbb{E}_X \mathbb{E}\left[Y_{X=\Delta(x,f)}(\{X = x\})\right] - \mathbb{E}[Y] > 0$.

For simplicity, the good incentives problem is stated as a decision problem. However, our subsequent results are unchanged if we instead consider related problems like producing a classifier that (optimally) incentivizes improvement. In the sequel, let `GoodIncentives` be an oracle for the Good Incentives problem. Then, `GoodIncentives` takes as input a cost function and a joint distribution over features and label, and returns whether or not a classifier with good incentives exists. We now turn to showing such an oracle is able to solve difficult causal inference problems.

5.2 A reduction from causal inference to designing good incentives

We now proceed to establish a reduction from causal inference to designing classifiers that incentive improvement. More specifically, we reduce from orienting the edges in a causal graph to the good incentives problem under a natural *manipulability* assumption. As a corollary, we prove this assumption holds in a broad family of causal graphs: additive noise models.

Most modern accounts of causality have manipulation or intervention at their core (Pearl, 2009; Spirtes et al., 2000; Woodward, 2003; Hoover, 2001). For example, Hoover (2001) insists that causality is defined by manipulation:

> The essence of causality is structure. Causal structures are mechanisms through which one thing is used to control or manipulate another.

The core idea of Hoover's account of causality is that if V causes W, then V can be manipulated to change W. Our main assumption asks for a similar linkage between structure and manipulation. Namely, for any edge $V \to W$ in a causal graph, we assume there exists some intervention on V that increases the expectation of W.

Assumption 5.1 (Manipulability)**.** *Let $G = (X, E)$ be a causal graph. Let X_{-W} denote the random variables X excluding node W, and similarly let x_{-w} be a realization of X_{-W}. For any edge $(V, W) \in E$ with $V \to W$, assume there exists v^*, possibly dependent on x_{-w}, such that*

$$\mathbb{E}_{X_{-W}} \mathbb{E}\left[W_{V := v^*(x_{-w})}(\{X_{-W} = x_{-w}\}) \right] > \mathbb{E}[W],$$

Importantly, the intervention v^* discussed in Assumption (5.1) is an intervention in the counterfactual model, *after observing the remaining variables $X \setminus \{W\}$* in the graph. In the strategic classification setting, this corresponds to choosing an adaptation conditional on the values of the observed features.

Assumption (5.1) is naturally satisfied in many causal models. Indeed, in Proposition (5.1), we prove assumption (5.1) holds for additive noise models when G is faithful. However, we first state and prove the main result of this section. Under Assumption (5.1), we exhibit a reduction from orienting the edges in a causal graph to the good incentives problem.

Theorem 5.1. *Let G be a causal graph with $|E|$ edges that satisfies Assumption* (5.1)*. Given the skeleton of G, using $|E|$ calls to* `GoodIncentives`*, we can orient all of the edges in G.*

Proof of Theorem (5.1). Let X denote all of the variables in the causal model, and let $X_i - X_j$ be an undirected edge in the skeleton G. We show how to orient $X_i - X_j$ with a single oracle call. Let $X_{-j} \triangleq X \setminus \left\{X_j\right\}$ be the set of features excluding X_j, and let x_{-j} denote an observation of X_{-j}.

Call `GoodIncentives` with features X_{-j}, label X_j, and the following cost function:

$$c(a; x_{-j}) = 2\mathbb{I}[a_k \neq 0 \text{ for any } k \neq i].$$

In other words, the individuals pays no cost to take an action that only affects X_i, but pays cost 2 if the action involves other variables. Since $f(x) \leq 1$ for any x, a best responding agent will therefore only take actions that affect X_i.

We now show `GoodIncentives` returns `Yes` if and only if $X_i \to X_j$. First, suppose $X_i \to X_j$. Then, by Assumption (5.1), there exists some $x_i^*(x_{-j})$, depending only on features x_{-j}, so that

$$\mathbb{E}_{X_{-j}}\mathbb{E}\Big[X_{j\left[X_i := x_i^*(x_{-j})\right]}\Big(\big\{X_{-j} = x_{-j}\big\}\Big)\Big] > \mathbb{E}\Big[X_j\Big].$$

This intervention $x_i^*(x_{-j})$ is incentivizable by the classifier

$$f(x_{-j}) = \mathbb{I}\Big[x_i = x_i^*(x_{-j})\Big].$$

In particular, for an individual with features x_{-j}, the action a where $a_i = x_i^*(x_{-j}) - x_i$ and otherwise $a_k = 0$ is the only best response to f. Therefore, the classifier f incentivizes improvement. Since some classifier with good incentivizes exists, the oracle `GoodIncentives` returns `Yes`.

On the other hand, suppose $X_i \leftarrow X_j$. Then no intervention on X_i can change the expectation of X_j. Clearly X_i is not a parent of X_j, and, moreover, X_i cannot be an ancestor of X_j since if there existed some indirect $X_i \cdots \to Z \cdots \to X_j$ path, then G would contain a cycle. Since no possible intervention exists, `GoodIncentives` must return `No`.

Repeating this procedure for each edge in the causal graph thus fully orients the skeleton with $|E|$ calls to `GoodIncentives`. □

We now turn to showing that Assumption (5.1) holds in a large class of nontrivial causal model, namely additive noise models (Peters et al., 2017).

Definition 5.2 (Additive Noise Model)**.** A structural causal model with graph $G = (X, E)$ is an additive noise model if the structural assignments are of the form

$$X_j := g_j(\mathbf{PA}_j) + U_j \quad \text{for } j = 1, \ldots, n\,.$$

Further, we assume that all nodes X_i are non-degenerate and that their joint distribution has a strictly positive density.[1]

Before proving the result, we need one additional technical assumption, namely faithfulness. The faithfulness assumption is ubiquitous in causal graph discovery setting and rules out additional conditional independence statements that are not implied by the graph structure. For more details and a precise statement of the d-separation criteria, see Pearl (2009).

Definition 5.3 (Faithful)**.** A distribution P_X is *faithful* to a DAG G if $A \perp\!\!\!\perp B \mid C$ implies that A and B are d-separated by C in G

Proposition 5.1. *Let $(X_1, \ldots, X_n)$ be an additive noise model, and let the joint distribution on $(X_1, \ldots, X_n)$ be faithful to the graph G. Then, G satisfies assumption* (5.1)*.*

Proof. Let $V \to W$ be an edge in G. We show there exists an intervention v^*, possibly dependent on the other nodes x_{-w}, that will increase the expected value of W. Therefore, we first condition on observing the remaining nodes $X_{-W} = x_{-w}$. In an additive noise model, given $X_{-W} = x_{-w}$ the

[1] The condition that the nodes X have a strictly positive density is met when, for example, the functional relationships f_i are differentiable and the noise variables U_i have a strictly positive density (Peters et al., 2017).

exogenous noise terms for all of the ancestors of W can be uniquely recovered. In particular, the noise terms are determined by

$$u_j = x_j - g_j(\mathbf{PA}_j).$$

Let $U_{\mathbf{A}}$ denote the collection of noise variables for ancestors of W *excluding* V. Both $U_{\mathbf{A}} = u_{\mathbf{A}}$ and $V = v$ are fixed by $X_{-W} = x_{-w}$.

Consider the structural equation for W, $W = g_W(\mathbf{PA}_W) + U_W$. The parents of W, $\mathbf{PA}_W$, are deterministic given V and $U_{\mathbf{A}}$. Therefore, given $V = v$ and $U_{\mathbf{A}} = u_{\mathbf{A}}$, $g_W(\mathbf{PA}_W)$ is a deterministic function of v and $u_{\mathbf{A}}$, which we write $h_W(v, u_{\mathbf{A}})$.

Now, we argue h_W is not constant in v. Suppose h_W were constant in v. Then, for every $u_{\mathbf{A}}$, $h_W(v, u_{\mathbf{A}}) = k(u_{\mathbf{A}})$. However, this means $W = k(U_{\mathbf{A}}) + U_W$, and $U_{\mathbf{A}}$ is independent of V, so we find that V and W are independent. However, since $V \to W$ in G, this contradicts faithfulness.

Since h_W is not constant in v, there exists at least one setting of $u_{\mathbf{A}}$ with v, v' so that $h_W(v', u_{\mathbf{A}}) > h_W(v, u_{\mathbf{A}})$. Since X has positive density, (v, u_a) occurs with positive probability. Consequently, if $v^*(u_{\mathbf{A}}) = \arg\max_v h_W(v, u_{\mathbf{A}})$, then

$$\begin{aligned}
\mathbb{E}_{X_{-W}}\mathbb{E}\left[W_{V:=v^*(u_{\mathbf{A}})}(\{X_{-W} = x_{-w}\})\right] &= \mathbb{E}_{X_{-W}}\left[\mathbb{E}[U_W] + \mathbb{E}[h_W(v^*(U_A), U_A) \mid X_{-W} = x_{-w}]\right] \\
&> \mathbb{E}[U_W] + \mathbb{E}_{X_{-W}}\mathbb{E}[h_W(V, U_A) \mid X_{-W} = x_{-w}] \\
&= \mathbb{E}[U_W] + \mathbb{E}[g_W(\mathbf{PA}_W)] \\
&= \mathbb{E}[W].
\end{aligned}$$

Finally, notice $v^*(u_{\mathbf{A}})$ can be computed solely form x_{-w} since $u_{\mathbf{A}}$ is fixed given x_{-w}. Together, this establishes that assumption (5.1) is satisfied for the additive noise model. □

On the other hand, assumption (5.1) can indeed fail in non-trivial cases.

Example 5.1. Consider a two variable graph with $X \to Y$. Let $Y = \varepsilon X$ where X and ε are independent and $\mathbb{E}[\varepsilon] = 0$. In general, X and Y are not independent, but for any x, $\mathbb{E}[Y_{X:=x}] = x\mathbb{E}[\varepsilon] = 0 = \mathbb{E}[Y]$.

6 Related Work

We have argued that causal inference is necessary for reasoning about the incentives produced by machine learning classifiers. We first explained how a causal perspective clarifies the difference between two viewpoints on adaptation: "improvement" and "gaming". Changes to causal features correspond to improvement and changes to non-causal features correspond to gaming. Although they do not explicitly use the language of causality, law scholars Bambauer and Zarsky (Bambauer and Zarsky, 2018) make a qualitatively equivalent characterization of when adaptation is gaming or not.

Where adaptation has been addressed in machine learning, it has often prioritized one viewpoint (gaming or improvement) at the expense of the other. For example, work in *strategic classification* typically assumes that all adaptation is gaming, and seeks to create classifiers that are *robust* to user adaptation (Brückner et al., 2012; Dalvi et al., 2004; Hardt et al., 2016). But since adaptation that is improvement is not distinguished from adaptation that is gaming, the resulting classifiers can make it harder for individuals to improve their classification outcome,

even when they are also improving their true target measure. These solution concepts can in turn lead to undesirable social burden (Hu et al., 2019; Milli et al., 2019).

Work on *counterfactual explanations* (Wachter et al., 2017) is often motivated as making it easier for individuals to improve their classification outcome. The work frames adaptation in a positive light and the improvement view is emphasized over the gaming view. However, imagine that one gives a counterfactual explanation of the form, "To reduce your insurance premium, you should stop posting photos of yourself smoking." It is clear that such an explanation is likely to lead to gaming rather than improvement. Thus, to avoid facilitating gaming, a causal perspective is important in determining the content of counterfactual explanations.

Motivated by goals similar to counterfactual explanations, Ustun et al. (2019) propose an audit measure known as *recourse*, defined as the ability of a person to obtain a desired outcome from a fixed model. They argue that recourse is important for giving users autonomy over their classification outcomes. For example, if someone is denied a loan, and the classifier has no recourse, then the person has no way to improve their classification, revealing a prominent lack of autonomy in the decision-making process. However, again, imagine a classifier whose outcomes can be changed on the basis of posting or not posting photos of smoking. This classifier has recourse, since, after all, you can change its classification outcome. However, the way you change the outcome is through gaming, and in particular, through self-censorship. And one typically wouldn't want to give a classifier "points" for inducing gaming incentives or self-censorship incentives. To this point, Ustun et al. (2019) acknowledge that allowing non-causal adaptations to factor into the measure of recourse may be undesirable. They note that the measure of recourse can be modified to exclude adaptations to non-causal features. However, in order to modify the measure one must have knowledge which features are causal and which are non-causal.

From the decision-maker's perspective, Khajehnejad et al. (2019) argue individual improvement is often aligned with institutional utility and treating all adaptation as gaming is overly pessimistic. This motivates a shift to reasoning about the decision maker's utility under strategic adaptation. In the strategic setting, they show maximizing institutional utility is NP-hard, and moreover optimal policies are generally stochastic. Consequently, even when the institution knows the causal model and can perfectly evaluate dynamic utility, finding policies that maximize dynamic utility is computationally challenging, and the resulting solutions are significantly more complex than in the non-strategic setting.

The creation of decision rules with optimal incentives has been long studied in economics, notably in principle-agent games (Ross, 1973; Grossman and Hart, 1992). Incentive design has recently been studied within a machine learning context. Kleinberg and Raghavan (2019) study the problem of producing a classifier that incentivizes a given "effort profile", the amount of desired effort an individual puts into certain actions. However, they assume "the evaluator has an opinion on which forms of agent effort they would like to promote." In other words, they assume the decision maker already knows which interventions lead to improvement.

As Theorem 5.1 makes clear, determining which interventions to incentivize requires causal reasoning. Consequently, causality enters in incentive design one step before the problem considered by Kleinberg and Raghavan (2019), which assumes a given, desired effort profile. Causality is needed to figure out what the desired effort profile is in the first place. For example, we think of doing homework and studying course material as good efforts to incentivize because we think they are causal to furthering one's knowledge.

In this paper, we have emphasized the role that causality plays in distinguishing between

what people consider to be "improvement" or "gaming". But what gets categorized as "improvement" or "gaming" also often reflects a moral judgement—gaming is bad, but improvement is good. And usually good or bad means good or bad from the perspective of the system operator. Ziewitz (2019) analyzes how adaptation comes to be seen as "ethical" or "unethical" through a case study on the practice of search engine optimization consultants. Using data from a large qualitative study of Twitter users, Burrell et al. (2019) argue that gaming can also be a form of individual "control" over the decision rule and that the exercise of control can be legitimate independently of whether an action is considered "gaming" or "improvement" in our framework.

Our causal framework focuses on the incentives of individuals being classified. Everitt et al. (2019) create a causal framework that focuses on related questions around the incentives of decision-makers. For example, their framework supports asking questions about which features the designer of a classifier has an incentive to use. This is particularly pertinent from an fairness perspective where we may want to know when the designer has an incentive to use protected attributes like race or gender in their classifier.

References

Jane Bambauer and Tal Zarsky. The algorithm game. *Notre Dame L. Rev.*, 94:1, 2018.

Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13(Sep):2617–2654, 2012.

Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. When users control the algorithms: values expressed in practices on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 3:19, 2019.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108. ACM, 2004.

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. ACM, 2018.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 178–184. AUAI Press, 2005.

Tom Everitt, Pedro A Ortega, Elizabeth Barnes, and Shane Legg. Understanding agent incentives using causal influence diagrams, part i: single action settings. *arXiv preprint arXiv:1902.09980*, 2019.

Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. In *Foundations of Insurance Economics*, pages 302–340. Springer, 1992.

DJ Hand, KJ McConway, and E Stanghellini. Graphical models of applicants for credit. *IMA Journal of Management Mathematics*, 8(2):143–155, 1997.

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. ACM, 2016.

Tatiana Homonoff, Rourke O'Brien, and Abigail B Sussman. Does knowing your fico score change financial behavior? evidence from a field experiment with student loan borrowers. Working Paper 26048, National Bureau of Economic Research, July 2019. URL http://www.nber.org/papers/w26048.

Kevin D Hoover. *Causality in Macroeconomics*. Cambridge University Press, 2001.

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

Moein Khajehnejad, Behzad Tabibian, Bernhard Schölkopf, Adish Singla, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844. ACM, 2019.

Marios Kokkodis, Panagiotis Papadimitriou, and Panagiotis G Ipeirotis. Hiring behavior models for online labor markets. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 223–232. ACM, 2015.

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. ACM, 2019.

Stephanie O'Neill. As insurers offer discounts for fitness trackers, wearers should step with caution, Nov 2018. URL https://www.npr.org/sections/health-shots/2018/11/19/668266197/as-insurers-offer-discounts-for-fitness-trackers-wearers-should-step-with-caution.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. MIT Press, 2017.

Stephen A Ross. The economic theory of agency: The principal's problem. *The American Economic Review*, 63(2):134–139, 1973.

Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

Marilyn Strathern. Improving ratings: audit in the british university system. *European review*, 5(3):305–321, 1997.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016.

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. ACM, 2019.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gpdr. *Harv. JL & Tech.*, 31:841, 2017.

James Woodward. *Making things happen: A theory of causal explanation*. Oxford University Press, 2003.

Malte Ziewitz. Rethinking gaming: The ethical work of optimization in web search engines. *Social studies of science*, page 0306312719865607, 2019.