Strategic Classification is Causal Modeling in Disguise

John Miller [1] **Smitha Milli** [1] **Moritz Hardt** [1]

Abstract

Consequential decision-making incentivizes individuals to strategically adapt their behavior to the specifics of the decision rule. While a long line of work has viewed strategic adaptation as gaming and attempted to mitigate its effects, recent work has instead sought to design classifiers that incentivize individuals to improve a desired quality. Key to both accounts is a cost function that dictates which adaptations are rational to undertake. In this work, we develop a causal framework for strategic adaptation. Our causal perspective clearly distinguishes between gaming and improvement and reveals an important obstacle to incentive design. We prove any procedure for designing classifiers that incentivize improvement must inevitably solve a non-trivial causal inference problem. We show a similar result holds for designing cost functions that satisfy the requirements of previous work. With the benefit of hindsight, our results show much of the prior work on strategic classification is causal modeling in disguise.

1. Introduction

Individuals faced with consequential decisions about them often use knowledge of the decision rule to strategically adapt towards achieving a desirable outcome. Much work in computer science views such *strategic adaptation* as adversarial behavior (Dalvi et al., 2004; Brückner et al., 2012), manipulation, or *gaming* (Hardt et al., 2016; Dong et al., 2018). More recent work rightfully recognizes that adaptation can also correspond to attempts at self-improvement (Bambauer & Zarsky, 2018; Kleinberg & Raghavan, 2019). Rather than seek classifiers that are robust to gaming, these works suggest to design classifiers that explicitly *incentive improvement* on some target measure (Kleinberg & Raghavan, 2019; Alon et al., 2020; Khajehnejad et al., 2019; Haghtalab et al., 2020).

Incentivizing improvement requires a clear distinction between gaming and improvement. While this distinction may be intuitive in some cases, in others, it is subtle. Do employer rewards for punctuality improve productivity? It sounds plausible, but empirical evidence suggests otherwise (Gubler et al., 2016). Indeed, the literature is replete with examples of failed incentive schemes (Oates & Schwab, 2015; Rich & Larson, 1984; Belot & Schröder, 2016).

Our contributions in this work are two-fold. First, we provide the missing formal distinction between gaming and improvement. This distinction is a corollary of a comprehensive causal framework for strategic adaptation that we develop. Second, we give a formal reason why incentive design is so difficult. Specifically, we prove any successful attempt to incentivize improvement must have solved a non-trivial causal inference problem along the way.

1.1. Causal Framework

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. We then characterize gaming and improvement by reasoning about how the corresponding intervention affects the predictor $\hat{Y}$ and the target variable Y. This is illustrated in Figure 1.

We combine the causal model with an *agent-model* that describes how individuals with a given setting of features respond to a classification rule. For example, it is common in strategic classification to model agents as being rational with respect to a *cost function* that quantifies the cost of feature changes.

Combining the causal model and agent model, we can separate improvement from gaming. Informally speaking, improvement corresponds to the case where the agent response to the predictor causes a positive change in the target variable Y. Gaming corresponds to the case where the agent response causes a change in the prediction $\hat{Y}$ but not the underlying target variable Y. Making this intuition precise, however, requires the language of counterfactuals of the

[1]Department of Computer Science, University of California, Berkeley, Berkeley, California, USA. Correspondence to: John Miller <miller_john@berkeley.edu>.

Proceedings of the 37^{th} *International Conference on Machine Learning*, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

form: What value would the variable Y have taken had the individual changed her features to X' given that her original features were X?

If we think of the predictor as a *treatment*, we can analogize our notion of improvement with the established causal quantity known as *effect of treatment on the treated*.

1.2. Inevitability of Causal Analysis

Viewed through this causal lens, only adaptations on causal variables can lead to improvement. Therefore, any mechanism for incentivizing improvement intuitively must capture some knowledge of the causal relationship between the features and the target measure. We formalize this intuition and prove causal modeling is unavoidable in incentive design. Specifically, we establish a computationally efficient reduction from discovering the causal structure relating the variables (sometimes called causal graph discovery) to a sequence of incentive design problems. In other words, designing classifiers to incentivize improvement is as hard as causal discovery. Previous work in strategic classification sidesteps this difficulty either by assuming the decision maker already has resolved this discovery step (Kleinberg & Raghavan, 2019; Alon et al., 2020), or by implicitly assuming all the features are causal for the label (Haghtalab et al., 2020).

Beyond incentivizing improvement, a number of recent works model individuals as acting in accordance with well-behaved cost functions that capture the difficulty of changing the target variable. We show constructing such *outcome-monotonic* cost functions also requires modeling the causal structure relating the variables, and we give a similar reduction from designing outcome-monotonic cost functions to causal discovery.

In conclusion, our contributions show that—with the benefit of hindsight—much work on strategic classification turns out to be causal modeling in disguise.

1.3. Related Work

This distinction between causal and non-causal manipulation in a classification setting is intuitive, and such considerations were present in early work on statistical risk assessment in lending (Hand et al., 1997). Although they do not explicitly use the language of causality, legal scholars Bambauer & Zarsky (2018) give a qualitatively equivalent distinction between gaming and improvement. While we focus on the incentives classification creates for individuals, Everitt et al. (2019) introduce a causal framework to study the incentives classification creates for decision-makers, e.g. which features the decision-maker is incentivized to use.

Numerous papers in strategic classification (Brückner et al., 2012; Dalvi et al., 2004; Hardt et al., 2016; Dong et al., 2018) focus on game-theoretic frameworks for preventing gaming. These frameworks form the basis of our agent-model, and Milli et al. (2019); Braverman & Garg (2020); Khajehnejad et al. (2019) introduce the outcome-monotonic cost functions we analyze in Section 5. Since these approaches do not typically distinguish between gaming and improvement, the resulting classifiers can be unduly conservative, which in turn can lead to undesirable social costs (Hu et al., 2019; Milli et al., 2019; Braverman & Garg, 2020).

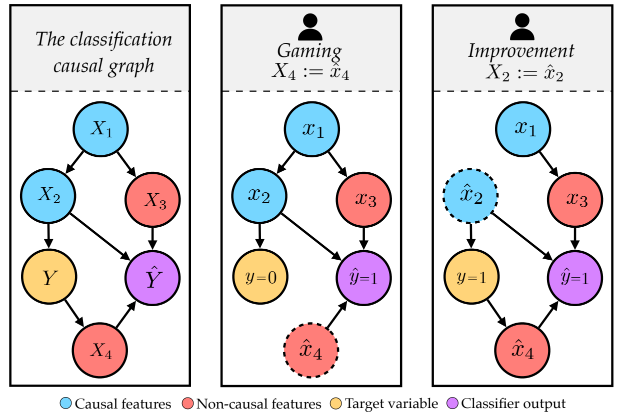

Figure 1. Illustration of the causal framework for strategic adaptation. Adaptation is modeled as interventions in a *counterfactual* causal graph, conditioned on the individual's initial features X. Gaming corresponds to interventions that change the classification $\hat{Y}$, but do not change the true label Y. Improvement corresponds to interventions that change both the classification $\hat{Y}$ and the true label Y. Incentivizing improvement requires inducing agents to intervene on *causal* features that can change the label Y rather than *non-causal* features. Distinguishing between these two categories of features in general requires causal analysis.

The creation of decision rules with optimal incentives, including incentives for improvement, has been long studied in economics, notably in principle-agent games (Ross, 1973; Grossman & Hart, 1992). In machine learning, recent work by Kleinberg & Raghavan (2019) and Alon et al. (2020) studies the problem of producing a classifier that incentivizes a given "effort profile", the amount of desired effort an individual puts into certain actions, and assumes the evaluator knows which forms of agent effort would lead to improvement, which is itself a form of causal knowledge. Haghtalab et al. (2020) seek to design classifiers that maximize improvement across the population, while Khajehnejad et al. (2019) seek to maximize institutional utility, taking into account both improvement and gaming. While these works do not use the language of causality, we demonstrate that these approaches nonetheless must perform some sort of causal modeling if they succeed in incentivizing improvement.

In this paper, we primarily consider questions of improvement or gaming from the perspective of the decision maker. However, what gets categorized as improvement or gaming also often reflects a moral judgement—gaming is bad, but improvement is good. 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 search engine optimization. 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.

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.

A 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. $X_i \in \mathbf{PA}_j$. We restrict ourselves to *Markovian* structural causal models, which have an acyclic causal graph and independent exogenous variables. The *skeleton* of a causal graph is the undirected version of the graph.

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 $X_i \coloneqq x_i$ that holds X_i at a fixed value. We use $\coloneqq$ 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 $X \coloneqq x$.

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 $X \coloneqq x$ 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 $X \coloneqq x$ where the distribution of exogenous variables U is updated by conditioning on the event E. We make heavy use of this *counterfactual* notion. For more details, see Pearl (2009).

3. A Causal Framework for Strategic Adaptation

In this section, we put forth a causal framework for reasoning about the incentives induced by a decision rule. Our framework consists of two components: *the agent model* and *the causal model*. The agent model is a standard component of work on strategic classification and determines what actions agents undertake in response to the decision rule. The causal model enables us to reason cogently about how these actions affect the agent's true label. Pairing these models together allow us to distinguish between incentivizing *gaming* and incentivizing *improvement*.

3.1. The Agent Model

As a running example, consider a software company that uses a classifier to filter software engineering job applicants. Suppose the model considers, among other factors, open-source contributions made by the candidate. Some individuals realize this and *adapt*—perhaps they polish their resume; perhaps they focus more of their energy on making open source contributions. The agent model describes precisely how individuals choose to adapt in response to a classifier.

As in prior work on strategic classification (Hardt et al., 2016; Dong et al., 2018), we model individuals as *best-responding* to the classifier. Formally, consider an individual with features $x \in \mathcal{X} \subseteq \mathbb{R}^n$, label $y \in \mathcal{Y} \subseteq \mathbb{R}$, and a classifier $f : \mathbb{R}^n \to \mathcal{Y}$. The individual has a set of available actions $\mathcal{A}$, and, in response to the classifier f, takes action $a \in \mathcal{A}$ to adapt her features from x to $x + a$. For instance, the features x might encode the candidate's existing open-source contributions, and the action a might correspond to making additional open-source contributions. Crucially, these modifications incur a *cost* $c(a; x)$, and the action the agent takes is determined by directly balancing the benefits of classification $f(x+a)$ with the cost of adaptation $c(a; x)$.

Definition 3.1 (Best-response agent model)**.** Given a cost function $c : \mathcal{A} \times \mathcal{X} \to \mathbf{R}_+$ and a classifier $f : \mathcal{X} \to \mathcal{Y}$, an individual with features x best responds to the classifier f by choosing action

$$a^* \in \arg\max_{a \in \mathcal{A}} f(x + a) - c(a; x).$$

Let $\Delta(x; f) = x + a^*$ denote a *best-response* of the agent to classifier f. When clear from context, we omit the dependence on f and write $\Delta(x)$.

In the best-response agent model, the cost function completely dictates what actions are rational for the agent to undertake and occupies a central modeling challenge. We discuss this further in Section 5. Our definition of the cost function in terms of an action set $\mathcal{A}$ is motivated by Ustun et al. (2019). However, this formulation is completely equivalent to the agent-models considered in other work (Hardt et al., 2016; Dong et al., 2018). In contrast to prior work, our main results only require that individuals approximately best-respond to the classifier.

Definition 3.2 (Approximate best-response)**.** For any $\varepsilon \in (0, 1)$, say $\Delta_\varepsilon(x, f) = x + \tilde{a}$ is an ε- best-response to classifier f if $f(x+\tilde{a}) - c(\tilde{a}; x) \geq \varepsilon \cdot (\max_a f(x+a) - c(a; x))$.

While we focus on a multiplicative approximation to the best-response, our results also hold for an additive approximation.

3.2. The Causal Model

While the agent model specifies which actions the agent takes in response to the classifier, the causal model describes how these actions effect the individual's true label.

Returning to the hiring example, suppose individuals decide increase their open-source contributions, X. Does this improve their software engineering skill, Y? There are two different causal graphs that explain this scenario. In one scenario, $Y \to X$: the more skilled one becomes, the more likely one is to contribute to open-source projects. In the other scenario, $X \to Y$: the more someone contributes to open source, the more skilled they become. Only in the second world, when $X \to Y$, do adaptations that increase open-source contributions raise the candidate's skill.

More formally, recall that a structural causal model has two types of nodes: endogenous nodes and exogenous nodes. In our model, the endogenous nodes are the individual's true label Y, their features $X = \{X_1, \ldots, X_n\}$, 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 the other unmodeled factors.

For an individual with features $X = x$, let $\Delta(x, f)$ denote the agent's response to classifier f. Since the agent chooses $\Delta(x, f)$ as a function of the observed features x, the label after adaptation is a *counterfactual* quantity. This, we model the individual's adaptation as an intervention in the submodel *conditioned on observing features* $X = x$. What value would the label Y take if the individual had features $\Delta(X, f)$, given that her features were originally X?

Formally, 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. Then, the label after adaptation is given by $Y_{X_A := \Delta(x,f)_A}(\{X = x\})$. 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, we omit reference to A and write $Y_{X:=\Delta(x,f)}(\{X = x\})$. In the language of potential outcomes, both X and Y are completely deterministic given the exogenous variables $U = u$, and we can express the label under adaptation as $Y_{X:=\Delta(x,f)}(u)$.

Much of the prior literature in strategic classification eschews explicit causal terminology and instead posits the existence of a "qualification function" or a "true binary classifier" $h : \mathcal{X} \to \mathcal{Y}$ that maps the individual's features to their "true quality" (Hardt et al., 2016; Hu et al., 2019; Braverman & Garg, 2020; Haghtalab et al., 2020). Such a qualification function should be thought of as the strongest possible causal model, where X is causal for Y, and the structural equation determining Y is completely deterministic.

3.3. Evaluating Incentives

Equipped with both the agent model and the causal model, we can formally characterize the incentives induced by a decision rule f. Key to our categorization is the notion of *improvement*, which captures how the classifier induces agents to change their label on average over the population baseline.

Definition 3.3. For a classifier f and a distribution over features X and label Y generated by a structural causal model, define the *improvement* incentivized by f, as

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

If $I(f) > 0$, we say that f *incentivizes improvement*. Otherwise, we say that f *incentivizes gaming*.

By the tower property, definition 3.3 can be equivalently written in terms of potential outcomes $I(f) = \mathbb{E}_U \left[Y_{X:=\Delta(x,f)}(U) - Y(U)\right]$. In this view, if we imagine exposure to the classifier f as a treatment, then improvement is the *treatment effect of exposure to classifier f on the label Y*. In general, since all individuals are exposed and adapt to the classifier in our model, and estimating improvement becomes an exercise in estimating the effect of treatment on the treated, and identifying assumptions are provided in Shpitser & Pearl (2009). Our notion of improvement is closely related to notion of "gain" discussed in Haghtalab et al. (2020), albeit with a causal interpretation.

While we focus on characterizing improvement at the population level for consistency with previous work, our framework easily accommodates other notions of improvement. For instance, we can similarly characterize improvement at the level of the individuals.

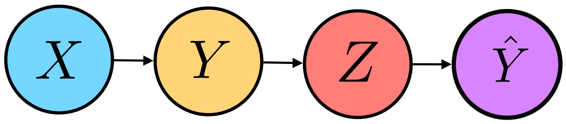

Figure 2. Reasoning about incentives requires both the agent-model and the causal model. The cost function plays a central role in the agent-model. Even though the classification $\hat{Y}$ only depends on the non-causal feature Z, the agent can change the label by manipulating, X, Z or both, depending on the cost function. The causal model determines how the agent's adaptation affects the target measure, but the agent model, and in turn the cost function, determines which actions the agent actually takes.

Definition 3.4. For a classifier f and a distribution over features X and label Y generated by a structural causal model, define the *improvement* incentivized by f for an individual with features x as

$$I(f;x) = \mathbb{E}\left[Y_{X:=\Delta(x,f)}(\{X=x\})\right] - \mathbb{E}\left[Y \mid X=x\right].$$

At first glance, the causal model and Definition 3.3 appear to offer a convenient heuristic for determining whether a classifier incentivizes gaming. Namely, does the classifier rely on non-causal features? However, even a classifier that uses purely non-causal features can still incentivize improvement if manipulating upstream, causal features is less costly than directly manipulating the non-causal features. The following example formalizes this intuition. Thus, reasoning about improvement requires considering both the agent model and the causal model.

Example 3.1. Suppose we have a structural causal model with features X, Z and label Y distributed as $X := U_X$, $Y := X + U_Y$, and $Z := Y + U_Z$, where $U_X, U_Y, U_Z \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Let the classifier f depend only on the non-causal feature, Z, $f(z) = \hat{y}$. Let $\mathcal{A} = \mathbb{R}^2$, and define the cost function $c(a;x) = (1/2)a^\top Ca$, where $C \succ 0$ is a symmetric, positive definite matrix with $\det(C) = 1$. Then, direct computation shows $\Delta(x,z;f) = (x - C_{12}, z + C_{11})$, and $I(f) = -C_{12}$. Hence, provided $C_{12} < 0$, f incentivizes improvement despite only rely on non-causal features. When $C_{12} < 0$ changing x and z jointly is less costly than manipulating z alone. This *complementarity* (Holmstrom & Milgrom, 1991) allows the decision-maker to incentivize improvement using only a non-causal feature. This example is illustrated in Figure 2.

4. Incentivizing Improvement Requires Causal Modeling

Beyond evaluating the incentives of a particular classifier, recent work has sought to *design* classifiers that explicitly incentivize improvement. Haghtalab et al. (2020) seeks classifiers that *maximize* the improvement of strategic individuals according to some quality score. Similarly, both Kleinberg & Raghavan (2019) and Alon et al. (2020) construct decision-rules that incentivize investment in a desired "effort profile" that ultimately leads to individual improvement. In this section, we show that when these approaches succeed in incentivizing improvement, they must also solve a non-trivial causal modeling problem. Therefore, while they may not explicitly discuss causality, much of this work is *necessarily* performing causal reasoning.

4.1. The Good Incentives Problem

We first formally state the problem of designing classifiers that incentivize improvement, which we call the *good incentives problem*. Consider the hiring example presented in Section 3. A decision-maker has knowledge of the joint distribution over the features (open-source contributions, coding test scores, etc) and the label (engineering ability), and wishes to design a decision rule that incentivizes strategic individuals to improve their engineering ability. As discussed in Section 3, the decision-maker must reason about the agent model governing adaptation, and we assume agent's *approximately* best-respond according to some specified cost function.

Definition 4.1 (Good Incentives Problem)**.** Assume agents ε-best-respond to the classifier for some $\varepsilon > 0$. Given:

1. A joint distribution $P_{X,Y}$ over examples $(x,y) \in \mathcal{X} \times \mathcal{Y}$ entailed by structural causal model, and
2. A cost function $c\colon \mathcal{A} \times \mathcal{X} \to \mathbb{R}_+$,

Find a classifier $f^*\colon \mathcal{X} \to \mathcal{Y}$ that incentivizes improvement, i.e. find a classifier with $I(f^*) > 0$. If no such classifier exists, output `Fail`.

The good incentives problem is closely related to the improvement problem studied in Haghtalab et al. (2020). Translated into our framework, Haghtalab et al. (2020) seek classifiers that optimally incentivize improvement and solve $\max_f I(f)$, which is a more difficult problem than finding *some* classier that leads to improvement.

In the sequel, let `GoodIncentives` be an oracle for the Good Incentives problem. `GoodIncentives` takes as input a cost function and a joint distribution over features and label, and either returns a classifier that incentivizes improvements or returns no such classifier exists.

4.2. A Reduction From Causal Modeling to Designing Good Incentives

Incentivizing improvement requires both (1) knowing which actions lead to improvement, and (2) incentivizing individuals to take those actions. Since only adaptation of causal features can affect the true label Y, determining which actions lead to improvement necessitates distinguishing between causal and non-causal features. Consequently, 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 establishes 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 discovery problem.

We prove this result under a natural assumption: improvement is always possible by manipulating causal features. In particular, for any edge $V \to W$ in the causal graph, there is always *some* intervention on V a strategic agent can take to improve W. We formally state this assumption below, and, as a corollary, we prove this assumption holds in a broad family of causal graphs: additive noise models.

Assumption 4.1. *Let $G = (X, E)$ be a causal graph, let X_{-W} denote the random variables X excluding node W. For any edge $(V, W) \in E$ with $V \to W$, there exists a real-valued function h mapping X_{-w} to an intervention $v^* = h(x_{-w})$ so that*

$$\mathbb{E}_{X_{-W}}\mathbb{E}\left[W_{V:=h(x_{-w})}\left(\{X_{-W} = x_{-w}\}\right)\right] > \mathbb{E}\left[W\right]. \quad (1)$$

Importantly, the intervention $v^* = h(x_{-w})$ discussed in Assumption 4.1 is an intervention in the counterfactual model, conditional on observing $X_{-W} = x_{-w}$. In strategic classification, this corresponds to choosing the adaptation conditional on the values of the observed features. Before proving Assumption 4.1 holds for faithful additive noise models, we first state and prove the main result.

Under Assumption 4.1, we exhibit a reduction from orienting the edges in a causal graph to the good incentives problem. While Assumption 4.1 requires Equation (1) to hold for every edge in the causal graph, it is straightforward to modify the result when Equation (1) only holds for a subset of the edges.

Theorem 4.1. *Let $G = (X, E)$ be a causal graph induced by a structural causal model that satisfies Assumption 4.1. Assume X has bounded support $\mathcal{X}$. Given the skeleton of G, using $|E|$ calls to* `GoodIncentives`*, we can orient all of the edges in G.*

Proof of Theorem 4.1. The reduction proceeds by invoking the good incentives oracle for each edge (X_i, X_j), taking X_j as the label and using a cost function that ensures only manipulations on X_i are possible for an ε-best-responding agent. If $X_i \to X_j$, then Assumption 4.1 ensures that improvement is possible, and we show `GoodIncentives` must return a classifier that incentivizes improvement. Otherwise, if $X_i \leftarrow X_j$, no intervention on X_i can change X_j, so `GoodIncentives` must return `Fail`.

More formally, 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 $X_{-j} \triangleq X \setminus \{X_j\}$ be the set of features excluding X_j, and let x_{-j} denote an observation of X_{-j}.

Consider the following good incentives problem instance. Let X_j be the label, and let the features be $(X_{-j}, \tilde{X}_i)$, where $\tilde{X}_i$ is an identical copy of X_i with structural equation $\tilde{X}_i := X_i$. Let the action set $\mathcal{A} = \mathbb{R}^n$, and let c be a cost function that ensures an ε-best-responding agent will only intervene on X_i. In particular, choose

$$c(a; (x_{-j}, \tilde{x}_i)) = 2B\mathbb{I}\left[a_k \neq 0 \text{ for any } k \neq i\right],$$

where $B = \sup\{\|x\|_\infty : x \in \mathcal{X}\}$. In other words, the individuals pays no cost to take actions that only affect X_i, but otherwise pays cost $2B$. Since every feasible classifier f takes values in $\mathcal{X}$, $f(x) \leq B$, and any action a with $a_k \neq 0$ leads to negative agent utility. At the same time, action $a = 0$ has non-negative utility, so an ε-best-responding agent can only take actions that affect X_i.

We now show `GoodIncentives` returns `Fail` if and only if $X_i \leftarrow X_j$. First, suppose $X_i \leftarrow X_j$. Then X_i is not a parent nor an ancestor of X_j since if there existed some $X_i \rightsquigarrow Z \rightsquigarrow X_j$ path, then G would contain a cycle. Therefore, no intervention on X_i can change the expectation of X_j, and consequently no classifier that can incentivize improvement exists, so `GoodIncentives` must return `Fail`.

On the other hand, suppose $X_i \to X_j$. We explicitly construct a classifier f that incentivizes improvement, so `GoodIncentives` cannot return `Fail`. By Assumption 4.1, there exists a function h so that

$$\mathbb{E}_{X_{-j}}\mathbb{E}\left[X_{j[X_i:=h(x_{-j})]}\left(\{X_{-j} = x_{-j}\}\right)\right] > \mathbb{E}\left[X_j\right].$$

Since $\tilde{X}_i := X_i$, Assumption 4.1 still holds additionally conditioning on $\tilde{X}_i = \tilde{x}_i$. Any classifier that induces agents with features $(x_{-j}, \tilde{x}_i)$ to respond by adapting only $X_i := h(x_{-j})$ will therefore incentivize improvement. The intervention $X_i := h(x_{-j})$ given $X_{-j} = x_{-j}$ is incentivizable by the classifier

$$f((x_{-j}, \tilde{x}_i)) = \mathbb{I}\left[x_i = h(\tilde{x}_{-j})\right],$$

where $\tilde{x}_j$ indicates that x_i is replaced by $\tilde{x}_i$ in the vector x_{-j}.

An ε-best-responding agent will choose action a^* where $a^*_i = h(\tilde{x}_{-j}) - x_i$ and otherwise $a^*_k = 0$ in response to

f. To see this, a^* has cost 0. Since $\tilde{X}_i := X_i$, we initially have $x_i = \tilde{x}_i$. Moreover, by construction, $h(\tilde{x}_{-j})$ depends only on the feature copy $\tilde{x}_i$, not x_i, so $h(\tilde{x}_{-j})$ is invariant to adaptations in x_i. Therefore, $h(\tilde{x}_{-j} + a^*_{-i}) = h(\tilde{x}_{-j}) = x_i + a^*_i$, so $f((x_{-j}, \tilde{x}_i) + a^*) = 1$. Thus, action a^* has individual utility 1, whereas all other actions have zero or negative utility, so any ε-best responding agent will choose a^*. Since all agents take a^*, it then follows by construction that $I(f) > 0$.

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 4.1 holds in a large class of nontrivial causal model, namely additive noise models (Peters et al., 2017).

Definition 4.2 (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 stating 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 4.3 (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 4.1. *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 4.1.*

The proof of Proposition 4.1 is deferred to the appendix.

On the other hand, Assumption 4.1 can indeed fail in nontrivial cases.

Example 4.1. Consider a two variable graph with $X \to Y$. Let $Y = UX$ where X and U are independent and $\mathbb{E}[U] = 0$. In general, X and Y are not independent, but for any x, x', $\mathbb{E}[Y_{X:=x'}(\{X = x\})] = x'\mathbb{E}[U] = 0 = \mathbb{E}[Y]$.

This section demonstrates the necessity of causal reasoning for incentivizing improvement. In the other direction, causal reasoning can be used to directly solve the good incentives problem. As discussed in Section 3, evaluating improvement corresponds to computing an effect of treatment on the treated. Abstractly, given an oracle for such queries, optimizing improvement becomes a generic stochastic optimization problem. Concretely, in subsequent work, Shavit et al. (2020) show how to leverage strategic response to evaluate causal interventions and give an efficient algorithm for using these interventions to design decision rules that incentivize improvement.

[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).

5. Designing Good Cost Functions Requires Causal Modeling

The cost function occupies a central role in the best-response agent model and essentially determines which actions the individual undertakes. Consequently, not few works in strategic classification model individuals as behaving according to cost functions with desirable properties, among which is a natural *monotonicity* condition—actions that raise an individual's underlying qualification are more expensive than those that do not. In this section, we prove an analogous result to the previous section and show constructing these cost functions also requires causal modeling.

5.1. Outcome-Monotonic Cost Functions

Although they use all slightly different language, Milli et al. (2019), Khajehnejad et al. (2019), and Braverman & Garg (2020) all assume the cost function is well-aligned with the label. Intuitively, they both assume (i) actions that lead to large increases in one's qualification are more costly than actions that lead to small increases, and (ii) actions that decrease or leave unchanged one's qualification have no cost. Braverman & Garg (2020) define these cost functions using an arbitrary qualification function that maps features X to label Y, while Milli et al. (2019) and Khajehnejad et al. (2019) instead use the *outcome-likelihood* $\Pr(y \mid x)$ as the qualification function. Khajehnejad et al. (2019) explicitly assume a causal factorization so that $\Pr(y \mid x)$ is invariant to interventions on X, and the qualification function of Braverman & Garg (2020) ensures a similar causal relationship between X and Y. Translating these assumptions into the causal framework introduced in Section 3, we obtain a class of *outcome-monotonic* cost functions.

Definition 5.1 (Outcome-monotonic cost)**.** A cost function $c : \mathcal{A} \times \mathcal{X} \to \mathbb{R}_+$ is *outcome-monotonic* if, for any features $x \in \mathcal{X}$:

1. For any action $a \in \mathcal{A}$, $c(a; x) = 0$ if and only if $\mathbb{E}[Y_{X:=x+a}(\{X = x\})] \leq \mathbb{E}[Y \mid X = x]$.

2. For pair of actions $a, a' \in \mathcal{A}$, $c(a; x) \leq c(a', x)$ if and

only if

$$\mathbb{E}\left[Y_{X:=x+a}(\{X=x\})\right] \leq \mathbb{E}\left[Y_{X:=x+a'}(\{X=x\})\right].$$

While several works assume the decision-maker has access to an outcome-monotonic cost, in general the decision-maker must explicitly construct such a cost function from data. This challenge results in the following problem.

Definition 5.2 (Learning outcome-monotonic cost problem)**.** Given action set $\mathcal{A}$ and a joint distribution $P_{X,Y}$ over a set of features X and label Y entailed by a structural causal model, construct an outcome-monotonic cost function c.

5.2. A Reduction From Causal Modeling to Constructing Outcome-Monotonic Costs

Outcome-monotonic costs are both conceptually desirable (Milli et al., 2019; Braverman & Garg, 2020) and algorithmically tractable (Khajehnejad et al., 2019). Simultaneously, outcome-monotonic cost functions encode significant causal information, and the main result of this section is a reduction from orienting the edges in a causal graph to learning outcome-monotonic cost functions under the same assumption as Section 4. Consequently, any procedure that can successfully construct outcome-monotonic cost functions must inevitably solve a non-trivial causal modeling problem.

Proposition 5.1. *Let $G = (X, E)$ induced by a structural causal model that satisfies Assumption 4.1. Let* `OutcomeMonotonicCost` *be an oracle for the outcome-monotonic cost learning problem. Given the skeleton of G, $|E|$ calls to* `OutcomeMonotonicCost` *suffices to orient all the edges in G.*

Proof. Let X denote the variables in the causal model, and let $X_i - X_j$ be an undirected edge. We can orient this edge with a single call to `OutcomeMonotonicCost`. Let $X_{-j} \triangleq X \setminus \{X_j\}$ denote the variables excluding X_j.

Construct an instance of the learning outcome-monotonic cost problem with features X_{-j}, label X_j, and action set $\mathcal{A} = \{\alpha e_i : \alpha \in \mathbb{R}\}$, where e_i is the i-th standard basis vector. In other words, the only possible actions are those that adjust the i-th coordinate. Let c denote the outcome-monotonic cost function returned by the oracle `OutcomeMonotonicCost`. We argue $c \equiv 0$ if and only if $X_i \leftarrow X_j$.

Similar to the proof of Theorem 4.1, if $X_i \leftarrow X_j$, then X_i can be neither a parent nor an ancestor of X_j. Therefore, conditional on $X_{-j} = x_{-j}$, there is no intervention on X_i that can change the conditional expectation of X_j. Since no agent has a feasible action that can increase the expected value of the label X_j and the cost function c is outcome-monotonic, c is identically 0.

On the other hand, suppose $X_i \to X_j$. Then, by Assumption 4.1, there is a real-valued function h such that

$$\mathbb{E}_{X_{-j}}\mathbb{E}\left[{X_j}_{X_i:=h(x_{-j})}(\{X_{-j}=x_{-j}\})\right] > \mathbb{E}\left[X_j\right].$$

This inequality along with the tower property then implies there is some agent x_{-j} such that

$$\mathbb{E}\left[{X_j}_{X_i:=h(x_{-j})}(\{X_{-j}=x_{-j}\})\right] > \mathbb{E}\left[X_j \mid X_{-j}=x_{-j}\right],$$

since otherwise the expectation would be zero or negative. Since $h(x_{-j})e_i \in \mathcal{A}$ by construction, there is some action $a \in \mathcal{A}$ that can increase the expectation of the label X_j for agents with features x_{-j}, so $c(a; x_{-j}) \neq 0$, as required. $\square$

The proof of Proposition 5.1 makes repeated calls to an oracle to construct outcome-monotonic cost functions to decode the causal structure of the graph G. In many cases, however, even a single outcome-monotonic cost function encode significant information about the underlying graph, as the following example shows.

Example 5.1. Consider a causal model with features (X, Z) and label Y with the following structural equations

$$\begin{aligned} X_i &:= U_{X_i} \quad \text{for } i = 1, \dots, n \\ Y &:= \sum_{i=1}^{n} \theta_i X_i + U_Y \\ Z_j &:= g_j(X, Y, U_{Z_j}) \quad \text{for } j = 1, \dots, m, \end{aligned}$$

for some set of non-zero coefficients $\theta_i \in \mathbb{R}$ and arbitrary functions g_j. In other words, the model consists of n causal features, m non-causal features, and a linear structural equation for Y.

Suppose the action set $\mathcal{A} = \mathbb{R}^{n+m}$, and let c be any outcome-monotonic cost. Then, $2(n+m)$ queries evaluations of c suffice to determine (1) which features are causal, and (2) $\text{sign}(\theta_i)$ for $i = 1, \dots, n$. To see this, evaluate the cost function at points $c(e_i; 0)$ and $c(-e_i; 0)$, where e_i denotes the i-th standard basis vector. Direct calculation shows

$$\mathbb{E}\left[Y_{(X,Z):=e_i}(\{(X,Z)=0\})\right] = \begin{cases} \theta_i & \text{if feature } i \text{ is causal} \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, since c is outcome-monotonic, if $c(e_i; 0) > 0$, then $\text{sign}(\theta_i) = 1$, if $c(-e_i; 0) > 0$, then $\text{sign}(\theta_i) = -1$, and if both $c(e_i; 0) = 0$ and $c(-e_i; 0) = 0$, then feature i is non-causal.

6. Discussion

The large collection of empirical examples of failed incentive schemes is a testament to the difficulty of designing incentives for individual improvement. In this work, we argued an important source of this difficulty is that incentivize design must inevitably grapple with causal analysis. Our results are not hardness or impossibility results per se. There are no fundamental computational or statistical barriers that prevent causal modeling beyond the standard unidentifiability results in causal inference. Indeed, subsequent work by Shavit et al. (2020) shows how causal queries can explicitly be leveraged to incentive improvement, and both Shavit et al. (2020) and Bechavod et al. (2020) prove strategic response itself can facilitate causal discovery. Our work suggests incentive design without causal understanding is unlikely to succeed, not that such understanding is unachievable.

Beyond incentive design, we hope our causal perspective clarifies intuitive, though subtle notions like gaming and improvement and provides a clear and consistent formalism for reasoning about strategic adaptation more broadly.

Acknowledgments

This research was generously supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1752814.

References

Alon, T., Dobson, M., Procaccia, A. D., Talgam-Cohen, I., and Tucker-Foltz, J. Multiagent evaluation mechanisms. In *AAAI*, pp. 1774–1781, 2020.

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

Bechavod, Y., Ligett, K., Wu, Z. S., and Ziani, J. Causal feature discovery through strategic modification. *arXiv preprint arXiv:2002.07024*, 2020.

Belot, M. and Schröder, M. The spillover effects of monitoring: A field experiment. *Management Science*, 62(1): 37–45, 2016.

Braverman, M. and Garg, S. The role of randomness and noise in strategic classification. *arXiv preprint arXiv:2005.08377*, 2020.

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

Burrell, J., Kahn, Z., Jonas, A., and Griffin, D. When users control the algorithms: values expressed in practices on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 3:19, 2019.

Dalvi, N., Domingos, P., Sanghai, S., Verma, D., et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108. ACM, 2004.

Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 55–70. ACM, 2018.

Eberhardt, F., Glymour, C., and Scheines, R. 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*, pp. 178–184. AUAI Press, 2005.

Everitt, T., Ortega, P. A., Barnes, E., and Legg, S. Understanding agent incentives using causal influence diagrams, part i: single action settings. *arXiv preprint arXiv:1902.09980*, 2019.

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

Gubler, T., Larkin, I., and Pierce, L. Motivational spillovers from awards: Crowding out in a multitasking environment. *Organization Science*, 27(2):286–303, 2016.

Haghtalab, N., Immorlica, N., Lucier, B., and Wang, J. Maximizing welfare with incentive-aware evaluation mechanisms. In *29th International Joint Conference on Artificial Intelligence*, 2020.

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

Hardt, M., Megiddo, N., Papadimitriou, C., and Wootters, M. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122. ACM, 2016.

Holmstrom, B. and Milgrom, P. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7(special_issue):24–52, 1991.

Hu, L., Immorlica, N., and Vaughan, J. W. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 259–268. ACM, 2019.

Khajehnejad, M., Tabibian, B., Schölkopf, B., Singla, A., and Gomez-Rodriguez, M. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.

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

Milli, S., Miller, J., Dragan, A. D., and Hardt, M. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 230–239. ACM, 2019.

Oates, W. E. and Schwab, R. M. The window tax: A case study in excess burden. *Journal of Economic Perspectives*, 29(1):163–80, 2015.

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

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

Rich, J. T. and Larson, J. A. Why some long-term incentives fail. *Compensation Review*, 16(1):26–37, 1984.

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

Shavit, Y., Edelman, B., and Axelrod, B. Learning from strategic agents: Accuracy, improvement, and causality. *arXiv preprint arXiv:2002.10066*, 2020.

Shpitser, I. and Pearl, J. Effects of treatment on the treated: identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 514–521, 2009.

Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19. ACM, 2019.

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

A. Missing Proofs

Proposition 4.1. Let $V \to W$ be an edge in G. We show there exists a real-valued function h that maps a realization of nodes $X_{-W} = x_{-w}$ to an intervention v^* that increases 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 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* those only have a path through 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 $\tilde{g}_W(v, u_{\mathbf{A}})$.

Now, we argue $\tilde{g}_W$ is not constant in v. Suppose $\tilde{g}_W$ were constant in v. Then, for every $u_{\mathbf{A}}$, $\tilde{g}_W(v, u_{\mathbf{A}}) = k(u_{\mathbf{A}})$. 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 $\tilde{g}_W$ is not constant in v, there exists at least one setting of $u_{\mathbf{A}}$ with v, v' so that $\tilde{g}_W(v', u_{\mathbf{A}}) > \tilde{g}_W(v, u_{\mathbf{A}})$. Since X has positive density, (v, u_a) occurs with positive probability. Consequently, if $h(u_{\mathbf{A}}) = \arg\max_v \tilde{g}_W(v, u_{\mathbf{A}})$, then

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

Finally, notice $h(u_{\mathbf{A}})$ can be computed solely form x_{-w} since $u_{\mathbf{A}}$ is fixed given x_{-w}. Together, this establishes that Assumption 4.1 is satisfied for the additive noise model. $\square$