The Social Cost of Strategic Classification

Smitha Milli John Miller Anca D. Dragan Moritz Hardt
University of California, Berkeley
{smilli,miller_john,anca,hardt}@berkeley.edu

Abstract

Consequential decision-making typically incentivizes individuals to behave strategically, tailoring their behavior to the specifics of the decision rule. A long line of work has therefore sought to counteract strategic behavior by designing more conservative decision boundaries in an effort to increase robustness to the effects of strategic covariate shift.

We show that these efforts benefit the institutional decision maker at the expense of the individuals being classified. Introducing a notion of social burden, we prove that any increase in institutional utility necessarily leads to a corresponding increase in social burden. Moreover, we show that the negative externalities of strategic classification can disproportionately harm disadvantaged groups in the population.

Our results highlight that strategy-robustness must be weighed against considerations of social welfare and fairness.

1 Introduction

As machine learning increasingly supports consequential decision making, its vulnerability to manipulation and gaming is of growing concern. When individuals learn to adapt their behavior to the specifics of a statistical decision rule, its original predictive power will deteriorate. This widely observed empirical phenomenon, known as Campbell's Law or Goodhart's Law, is often summarized as: "Once a measure becomes a target, it ceases to be a good measure" [25].

Institutions using machine learning to make high-stakes decisions naturally wish to make their classifiers robust to strategic behavior. A growing line of work has sought algorithms that achieve higher utility for the institution in settings where we anticipate a strategic response from the the classified individuals [10, 5, 14]. Broadly speaking, the resulting solution concepts correspond to more conservative decision boundaries that increase robustness to some form of covariate shift.

But there is a flip side to strategic classification. As insitutional utility increases as a result of more cautious decision rules, honest individuals worthy of a positive classification outcome may face a higher bar for success.

The costs incurred by individuals as a consequence of strategic classification are by no means hypothetical, as the example of lending shows. In the United States, credit scores are widely deployed to allocate credit. However, even creditworthy individuals routinely engage in artificial practices intended to improve their credit scores, such as opening up a certain number of credit lines in a certain time period [9]. In this work, we study the tension between institutional and individual utility in strategic classification. We first introduce a general measure of the cost of strategic classification, which we call the *social burden*. Informally, the social burden measures the expected cost that a positive individual needs to incur to be correctly classified as positive.

For a broad class of cost functions, we prove there exists an intrinsic trade-off between institutional accuracy and social burden: any increase in institutional accuracy comes at an increase in social burden. Moreover, we precisely characterize this trade-off and show the commonly considered Stackelberg equilibrium solution that achieves maximal institutional accuracy comes at the expense of maximal social burden.

Equipped with this generic trade-off result, we turn towards a more careful study of how the social burden of strategic classification impacts different subpopulations. We find that the social burden can fall disproportionally on disadvantaged subpopulations, under two different notions by which one group can be *disadvantaged* relative to another group. Furthermore, we show that as the institution improves its accuracy, it exacerbates the gap between the burden to an advantaged and disadvantaged group. Finally, we illustrate these conditions and their consequences with a case study on FICO data.

1.1 Our Contributions

In this paper, we make the following contributions:

1. We prove a general result demonstrating the trade-off between institutional accuracy and individual utility in the strategic setting. Our theoretical characterization is supplemented with examples illustrating when an institution would prefer to operate along different points in this trade-off curve.

2. We show fairness considerations inevitably arise in the strategic setting. When individuals incur cost as a consequence of making a classifier robust to strategic behavior, we show the costs can disproportionally fall by disadvantaged subpopulations. Furthermore, as the institution increases its robustness, it also increases the disparity between the subpopulations.

3. Using FICO credit data as a case-study, we empirically validate our modeling assumptions and illustrate both the general trade-offs and fairness concerns involved with strategic classification in a concrete setting.

Reflecting on our results, we argue that the existing view of strategic classification has been *instituition-centric*, ignoring the social burden resulting from improved institutional utility. Our framework makes it possible to select context-specific trade-offs between institutional and individual utility, leading to a richer space of solutions.

Another key insight is that discussions of strategy-robustness must go hand in hand with considerations of fairness and the real possibility that robustness-promoting mechanisms can have disparate impact in different segments of the population.

2 Model

Strategic classification. Throughout this work, we consider the binary classification setting. Each individual has features $x \in \mathcal{X}$ and a label $y \in \mathcal{Y} = \{0, 1\}$. The *institution* publishes a classifier $f : \mathcal{X} \to \mathcal{Y}$. In the non-strategic setting, the *institutional utility* is simply the classification accuracy of f:

$$\mathcal{U}(f) = \mathbb{P}(f(x) = y)$$

In the *strategic setting*, the *individual* can modify their features x to new features x'. However, modification incurs a cost given by $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$. The *individual utility* after changing from x to x' is then $f(x') - c(x, x')$. We assume the individual optimally adapts their features to maximize this utility. The *best-response* of individual with features x to classifier f is given by

$$\Delta(x; f) = \arg\max_{x'} f(x') - c(x, x').$$

When it is clear from context we will drop the dependence on f and write the individual's best response as $\Delta(x)$. The expression above may not have a unique maximizer. We assume that the individual x does not adapt her features if she is already accepted by the classifier, i.e $f(x) = 1$, or if there is no maximizer x' she can move to such that $f(x') = 1$. In cases where the individual does adapt, we let x' be an arbitrary maximizer such that $f(x') = 1$.

In line with prior work [15], we assume that the institution has knowledge of the cost function c, although in practice, the cost function would likely need to be learned from data. When the institution knows the cost function, it can take into account how individuals will adapt when choosing what classifier to use. For example, imagine that the institution is trying to rank pages on a social network. Although the number of likes a page has may be predictive, if the institution knows that it is low cost for individuals to game how many likes they have, it can choose to weigh the feature less.

In the strategic setting, the institution's utility is modified to account for this manipulation of features. The *strategic utility* for the institution measures accuracy *after* individual responses:

$$\mathcal{U}_\Delta(f) = \mathbb{P}(f(\Delta(x)) = y).$$

Social burden. Focusing purely on maximizing $\mathcal{U}_\Delta$, as done in prior work [4, 15, 12], ignores the cost that a classifier imposes on individuals, particularly true positives. To measure this cost, we introduce the *social burden*, defined as the expected minimum cost $m_f(x) = \min_{f(x')=1} c(x, x')$ that positive individuals must incur in order to be classified correctly.

Definition 2.1 (Social burden)**.** The social burden of a classifier f is defined as

$$\mathcal{B}_+(f) = \mathbb{E}\left[\min_{f(x')=1} c(x, x') \mid y = 1\right].$$

The social burden measures two types of negative effects on positive individuals, depending on whether they change their features or not. Since individuals respond optimally, if $m_f(x)$, the minimum cost necessary to be accepted, is less than one, then the individual adapts their features.

On the other hand, if $m_f(x)$ is greater than or equal to one, then the individual does not adapt their features because the cost of changing their features outweighs the benefit she gets from being accepted.

In the first case, the individual still gets accepted, but incurs a cost for changing their features. In the second case, the individual does not adapt their features and does not get accepted by the classifier, so it is more appropriate to view $m_f(x)$ as a hypothetical cost that blocks the individual from being accepted. The social burden, which takes an expectation over $m_f(x)$, measures both the cost incurred in order to be accepted and the hypothetical cost that prevents acceptance.

Assumptions on cost function. While there are many possible models for the cost function, we restrict our attention to a natural set of cost functions that we call *outcome monotonic*. Outcome monotonic costs capture two intuitive properties: (1) Monotonically improving one's outcome requires monotonically increasing amounts of work, and (2) it is zero cost to worsen one's outcome. This captures the intuition that, for example, it is harder to pay back loans than it is to go bankrupt.

Definition 2.2 (Outcome likelihood)**.** The outcome likelihood of an individual x is $\ell(x) = \mathbb{P}(Y = 1 \mid X = x)$.

Definition 2.3 (Outcome Monotonic Cost)**.** A cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is outcome monotonic if for any $x, x', x^* \in X$ the following properties hold.

- *Zero-cost to move to lower outcome likelihoods.* $c(x, x') > 0$ if and only if $\ell(x) > \ell(x')$.
- *Monotonicity in first argument.* $c(x, x^*) > c(x, x') > 0$ if and only if $\ell(x^*) > \ell(x') > \ell(x)$.
- *Monotonicity in second argument.* $c(x, x^*) > c(x, x') > 0$ if and only if $\ell(x^*) > \ell(x') > \ell(x)$.

Under these assumptions, we can equivalently express the cost as a cost over outcome likelihoods, $c_L : \ell(\mathcal{X}) \times \ell(\mathcal{X}) \to \mathbb{R}_{\geq 0}$, defined in the following lemma.

Lemma 2.1. *When the cost function $c(x, x')$ is outcome monotonic, then it can be written as a cost function over outcome likelihoods $c_L(l, l') \coloneqq c(x, x')$ where $x, x' \in \mathcal{X}$ are any points such that $l = \ell(x)$ and $l' = \ell(x')$.*

Proof. The monotonicity assumptions imply that if $\ell(x^*) = \ell(x')$, then $c(\cdot, x') = c(\cdot, x^*)$ and $c(x', \cdot) = c(x^*, \cdot)$. Thus, $c_L(l, l') \coloneqq c(x, x')$ is well-defined because any points x and x' such that $l = \ell(x)$ and $l' = \ell(x')$ yireld the same value of $c(x, x')$. □

Throughout the paper, we will make use of the equivalent likelihood cost c_L when a proof is more naturally expressed with c_L, rather than with the underlying cost c.

3 Institutional Utility Versus Social Burden

In this section, we characterize the inherent trade-offs between institutional utility and social burden in the strategic setting. In particular, we show any classifier that improves institutional utility over the best classifier in the static setting causes a corresponding increase in social burden.

We prove this result in two steps. First, we prove any classifier can be represented as a threshold classifier that depends only on a threshold $\tau \in [0, 1]$. Then, we show increasing utility for the institution corresponds to raising this threshold τ, but that the social burden monotonically increases in τ.

Equipped with this result, we show the (Pareto-optimal) set of classifiers that increase institutional utility in the strategic setting corresponds to an interval I. Each threshold $\tau \in I$ represents a particular trade-off between institutional utility and social burden. Strategic classification corresponds to one extremum: the best strategic utility but the worst social burden. The non-strategic utility corresponds to the other: doing nothing to prevent gaming. Neither is likely to be the right trade-off in practical contexts. Real domains will require a careful weighting of these two utilities, leading to a choice somewhere in between. Thus, a main contribution of our work is exposing this interval.

3.1 General Trade-Off

We now proceed to prove the trade-off between institutional utility and social burden. Our first step is to show that in the strategic setting we can restrict attention to classifiers that threshold on the outcome likelihood (assuming the cost is outcome monotonic as in Definition 2.3).

Definition 3.1 (Outcome threshold classifier)**.** An outcome threshold classifier f is a classifier of the form $f(x) = \mathbb{I}\{\ell(x) \geq \tau\}$ for $\tau \in [0, 1]$.

In practice, the institution may not know the outcome likelihood $\ell(x) = \mathbb{P}(Y = 1 \mid X = x)$. However, as shown in the next lemma, for any classifier that they do use, there is a threshold classifier with equivalent institutional utility and social burden. Thus, we can restrict our theoretical analysis to only consider threshold classifiers.

Lemma 3.1. *For any classifier f there is an outcome threshold classifier f' such that $\mathcal{U}_\Delta(f) = \mathcal{U}_\Delta(f')$ and $\mathcal{B}_+(f) = \mathcal{B}_+(f')$.*

Proof. Let $\tau(f) = \min_{x: f(x)=1} \ell(x)$ be the outcome likelihood at which an individual is accepted by the classifier f. Then, let $f' = \mathbb{I}\{\ell(x) \geq \tau(f)\}$ be the outcome threshold classifier that accepts all individuals above $\tau(f)$. We will show that the institutional utility and social burden of f and f' are equal.

Since the cost function is outcome monotonic, it is the same cost to move to any point with the same outcome likelihood. Furthermore, it is higher cost to move to points of higher likelihood, i.e, if $\ell(x') > \ell(x^*)$, then $c(x, x') > c(x, x^*)$. Since individuals game optimally, when an individual changes her features in response to the classifier f, she has no incentive to move to a point with probability higher than $\tau(f)$ – that would just cost more. Therefore, she will move to any point with likelihood $\tau(f)$ to be accepted by f and will incur the same cost, regardless of which point it is. Thus, we can write the set of individuals accepted by f, $\mathcal{A}_\Delta(f)$, as

$$\begin{aligned} \mathcal{A}_\Delta(f) &= \{x \mid f(\Delta(x)) = 1\} \\ &= \{x \mid \exists x' : f(x') = 1,\ c(x, x') \leq 1\} \\ &= \{x \mid c(x, x') \leq 1,\ x' : \ell(x) = \tau(f)\}\,. \end{aligned}$$

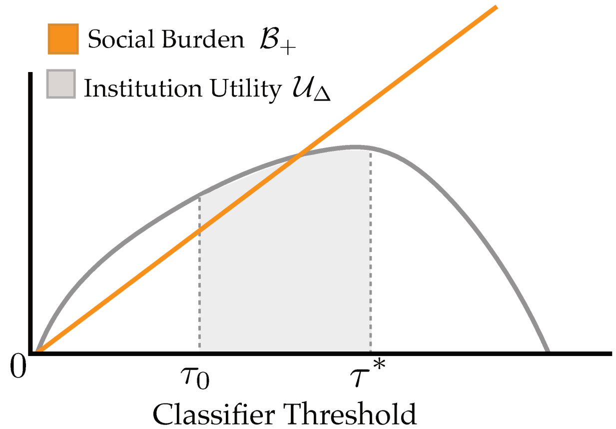


Figure 1: The general shapes of the institution utility and social burden as a function of the threshold the institution chooses. The threshold t_0 is the non-strategic optimal, while the threshold τ^* is the Stackelberg equilibrium.

Since $\tau(f) = \tau(f')$, the individuals accepted by f and f' are equal: $\mathcal{A}_\Delta(f) = \mathcal{A}_\Delta(f')$. Therefore, their institutional utilities $\mathcal{U}_\Delta(f)$ and $\mathcal{U}_\Delta(f')$ are equal. We can similarly show that the social burdens of f and f' are also equal:

$$\begin{aligned}
\mathcal{B}_+(f) &= \mathbb{E}\left[\min_{f(x')=1} c(x,x') \mid y=1\right], \\
&= \mathbb{E}\left[c(x,x') \mid y=1\right] \quad x' : \ell(x') = \tau(f) \\
&= \mathbb{E}\left[c(x,x') \mid y=1\right] \quad x' : \ell(x') = \tau(f') \\
&= \mathbb{E}\left[\min_{f'(x')=1} c(x,x') \mid y=1\right] = \mathcal{B}_+(f').
\end{aligned}$$

□

Since outcome threshold classifiers can represent all classifiers in the strategic setting, we will henceforth only consider outcome threshold classifiers. We will overload notation and use $\mathcal{U}_\Delta(\tau)$ and $\mathcal{B}_+(\tau)$ refer to $\mathcal{U}_\Delta(f_\tau)$ and $\mathcal{B}_+(f_\tau)$ where $f_\tau(x) = \mathbb{I}\{\ell(x) \geq \tau\}$ is the outcome threshold classifier with threshold τ.

Figure 1 illustrates how institutional utility and social burden change as the threshold of the classifier increases. The institutional utility is quasiconcave, while the social burden is monotonically non-decreasing. The next lemma provides a formal characterization of the shapes shown in Figure 1.

Theorem 3.1. *The institution utility $\mathcal{U}_\Delta(\tau)$ is quasiconcave in τ and has a maximum at a threshold $\tau^* \geq \tau_0$ where $\tau_0 = 0.5$ is the threshold of the non-strategic optimal classifier. The social burden $\mathcal{B}_+(\tau)$ is monotonically non-decreasing in τ. Furthermore, if $\mathcal{U}_\Delta(\tau) \neq \mathcal{U}_\Delta(\tau')$, then $\mathcal{B}_+(\tau) \neq \mathcal{B}_+(\tau')$.*

Proof. Let $\mathcal{A}_\Delta(\tau)$ and $\mathcal{A}(\tau)$ be the set of individuals accepted by f in the strategic and non-strategic setting, respectively. If $\tau < \tau_0$, we have $\mathcal{A}_\Delta(\tau) \supseteq \mathcal{A}_\Delta(\tau_0) \supseteq \mathcal{A}(\tau_0)$. Since $\mathcal{A}(\tau_0)$ is the optimal acceptance region, $\mathcal{U}_\Delta(\tau) \leq \mathcal{U}_\Delta(\tau_0)$. Therefore, if a threshold τ^* is optimal for the institution, i.e, $\mathcal{U}_\Delta(\tau^*) = \max_\tau \mathcal{U}_\Delta(\tau)$, then $\tau^* \geq \tau_0$.

Recall that a univariate function $f(z)$ is quasiconcave if there exists z^* such that f is non-decreasing for $z < z^*$ and is non-increasing for $z > z^*$. Note that if $\tau < \tau^*$ we have that $\mathcal{A}_\Delta(\tau) \supseteq \mathcal{A}_\Delta(\tau^*)$, thus $\mathcal{U}_\Delta(\tau) \le \mathcal{U}_\Delta(\tau^*)$. Similarly, for any $\tau > \tau^*$ we have that $\mathcal{A}_\Delta(\tau^*) \supseteq \mathcal{A}_\Delta(\tau)$, and thus $\mathcal{U}_\Delta(\tau) \le \mathcal{U}_\Delta(\tau^*)$. Therefore, $\mathcal{U}_\Delta(\tau)$ is quasiconcave in τ.

Let $c_x(\tau)$ be the cost required for a specific individual x to be classified positively: $c_x(\tau) = c(x, x')$ where x' is any point such that $\ell(x') = \tau$. The social burden can then be expressed as $\mathcal{B}_+(\tau) = \mathbb{E}[c_x(\tau) \mid y = 1]$. Since $c_x(\tau)$ is monotonically non-decreasing, $\mathcal{B}_+(\tau)$ is also monotonically non-decreasing.

Suppose $\mathcal{U}_\Delta(\tau) \neq \mathcal{U}_\Delta(\tau')$ and without loss of generality let $\tau < \tau'$. For all individuals x, $c_x(\tau') \ge c_x(\tau)$. If there is at least one individual x such that $c_x(\tau') > c_x(\tau)$, then $\mathcal{B}_+(\tau') > \mathcal{B}_+(\tau)$. But since $\mathcal{U}_\Delta(\tau) \neq \mathcal{U}_\Delta(\tau')$, there must exist an individual x such that $x \in \mathcal{A}_\Delta(\tau)/\mathcal{A}_\Delta(\tau')$ and $p(X = x \mid Y = 1) > 0$. For this individual $c_x(\tau') > c_x(\tau)$. Therefore, $\mathcal{U}_\Delta(\tau) \neq \mathcal{U}_\Delta(\tau') \implies \mathcal{B}_+(\tau') \neq \mathcal{B}_+(\tau)$. □

As a corollary, if the institution increases its utility beyond that attainable by the optimal classifier in the non-strategic setting, then the institution also causes higher social burden.

Corollary 3.1. *Let τ be any threshold and $\tau_0 = 0.5$ be the optimal threshold in the non-strategic setting. If $\mathcal{U}_\Delta(\tau) > \mathcal{U}_\Delta(\tau_0)$, then $\mathcal{U}_S(\tau) < \mathcal{U}_S(\tau_0)$.*

3.2 Choosing a Concrete Trade-off

We have now shown that increases in institutional utility come at a cost in terms of social burden and vice-versa. This still leaves open the question: what is the concrete trade-off an institution should choose?

Theorem 3.1 provides a precise characterization of the choices available to trade-off between institutional utility and social burden. The baseline choice for the institution is to not account for strategic behavior and use the non-strategic optimum τ_0. Maximizing utility without regard to social burden leads the institution to choose τ^*. In general, the interval $[\tau_0, \tau^*]$ offers the set of trade-offs the institution considers. Choosing $\tau > \tau_0$ can increase robustness at the price of increasing social burden. Thresholds $\tau > \tau^*$ are not Pareto-efficient and are not considered.

Much of the prior work in machine learning has focused exclusively on solutions corresponding to the thresholds at the extreme: τ_0 and τ^*. The threshold τ_0 is the solution when strategic behavior is not accounted for. The threshold τ^* is also known as the *Stackelberg equilibrium* and is the subject of recent work in strategic classification [4, 14, 12]. While using τ^* may be warranted in some cases, a proper accounting of social burden would lead institutions to choose classifiers somewhere *between* the extremes of τ_0 and τ^*.

The exact choice of $\tau \in [\tau_0, \tau^*]$ is *context-dependent* and depends on balancing concerns between institutional and broader social interest. We now highlight cases where using τ_0 or τ^* may be suboptimal, and using a threshold $\tau \in (\tau_0, \tau^*)$ that balances robustness with social burden is preferable.

Example 3.1 (Expensive features.)**.** If measuring a feature is costly for individuals and offers limited institutional gains, an institution may choose to ignore the feature, even if it means giving up accuracy on the margin. In an educational context, a university may decide to no longer require applicants to submit standardized test scores, which can cost applicants hundreds of dollars, if the corresponding improvement in admissions outcomes is very small [1].

Example 3.2 (Reducing social burden under resource constraints.)**.** Aid organizations use machine learning to determine where to allocate resources after natural disasters [18]. In these cases, positive individuals are precisely those people who are in need of aid and may experience very high costs to change their features. Using thresholds with high social burden is therefore undesirable. At the same time, organizations giving out aid often face significant resource constraints. False positives from individuals gaming the classifier ties up resources that could be better used elsewhere. Consequently, using the non-strategic threshold is also undesirable. The aid organization should choose a some threshold τ with $\tau_0 < \tau < \tau^*$ that reflects these trade-offs.

Example 3.3 (Misspecification of agent model.)**.** Strategic classification models (including ours) typically assume that the individual optimally responds to the classifier f. However, in reality, individuals will not have perfect knowledge of the classifier f when it is first deployed. Instead, they may be able to learn about how the classifier works over time, and gradually improve their ability to game the classifier. For example, self-published romance authors exchanged information in private chat groups about how to best game Amazon's book recommendation algorithms [19]. For the institution, it is difficult to a priori model the dynamics of how information about the classifier propagates. A preferable solution may be to simply make the assumption that the individual can best respond to the classifier, but to only gradually increase the threshold from the non-strategic τ_0 to the Stackelberg optimal τ^* over time.

In fact, misspecification of the agent model (described above), is why [5] suggest that the Stackelberg equilibrium is too conservative, and instead prefer to use Nash equilibrium strategies. Complementary to their observation, we show that there is a more general reason Nash equilibria may be preferable. Namely, that Nash equilibria have lower social burden than the Stackelberg solution. As the following lemma shows, in our context, the set of Nash equilibrium form an interval $[t_N, \tau^*] \subset I$ for some $t_N \geq t_0$. The proof is deferred to the appendix.

Lemma 3.2. *Suppose the cost over likelihoods c_L is continuous and $\ell(\mathcal{X}) = [0, 1]$, i.e, all likelihoods have non-zero support. Then, the set of Nash equilibrium strategies for the institution is $[\tau_N, \tau^*]$ for some $\tau_N \geq \tau_0$ where $\tau_0 = 0.5$ is the non-strategic optimal threshold and τ^* is the Stackelberg equilibrium strategy.*

The Stackelberg equilibrium requires the institution to choose τ^*, whereas Nash equilibria give the institution latitude to trade-off between institutional utility and social burden by choosing from the interval $[t_N, \tau^*] \subset I$. This provides an additional argument in favor of Nash equilibria– institutions can reason in terms of equilibria *and* achieve more favorable outcomes in terms of social burden.

4 Fairness to Subpopulations

Our previous section showed that increased robustness in the face of strategic behavior comes at the price of additional social burden. In this section, we show this social burden is not fairly distributed: when the individuals being classified are from latent subpopulations, say of race, gender, or socioeconomic status, the social burden can disproportionately fall on disadvantaged subpopulations. Furthermore, we find that improving the institution's utility can exacerbate the gap between the social burden incurred by an advantaged and disadvantaged group.

Concretely, suppose each individual is from a subpopulation $G \in \{a, b\}$. The social burden a classifier f has on a group $g \in G$ is the expected minimum cost required for a positive individual from group g to be accepted: $\mathcal{B}_{+,g}(f) = \mathbb{E}\left[\min_{f(x')=1} c(x, x') \mid Y = 1, G = g\right]$. We can then define the social gap between groups a and b:

Definition 4.1 (Social gap)**.** The *social gap* $\mathcal{G}(f)$ induced by a classifier f is the difference in the social burden to group b compared to a: $\mathcal{G}(f) = \mathcal{B}_{+,b}(f) - \mathcal{B}_{+,a}(f)$.

The social gap is a measure of how much more costly it is for a positive individual from group b to be accepted by the classifier than a positive individual from group a. For example, there is evidence that women need to attain higher educational qualifications than their male counterparts to receive the same salary [6].

A high social gap is alarming for two reasons. First, even when two people from group a and group b are equally qualified, the individual from group a may choose not to participate at all because of the cost she would need to endure to be accepted. Secondly, if she does decide to participate, she may continue to be at a disadvantage after being accepted because of the additional cost she had to endure, e.g., repaying student loans.

Non-strategic classification can already induce a social gap between two groups, and strategic classification can exacerbate this gap. We show this under two natural ways group b may be disadvantaged. In the first setting, the feature distributions of group a and b are such that a positive individual from group b is less likely to be considered positive, compared to group a. In the second setting, individuals from group b have a higher cost to adapt their features compared to group a. Under both of these conditions, any improvement the institution can make to its own strategic utility has the side effect of worsening (increasing) the social gap.

4.1 Different Feature Distributions

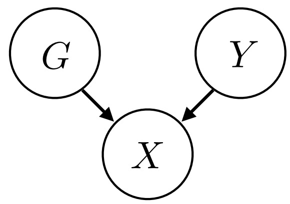
In the first setting we analyze, the way groups a and b differ is through their distributions over features. We say that group b is disadvantaged if the features distributions are such that positive individuals from group b are less likely to be considered positive than those from group a. Formally, this can be characterized as the following:

Definition 4.2 (Disadvantaged in features)**.** Let $L_{+,g} = \ell(X) \mid Y = 1, G = g$ be the outcome likelihood of a positive individual from group g, and let $F_{+,g}$ be the cumulative distribution function of $L_{+,g}$. We say that group b is disadvantaged in features if $F_{+,b}(l) > F_{+,a}(l)$ for all $l \in (0, 1)$.

In the economics literature, the relationship between $L_{+,a}$ and $L_{+,b}$ is referred to as *strict first-order stochastic dominance* [21]. An equivalent way to understand the definition is that group b is disadvantaged in features if and only if the distribution of $L_{+,a}$ can be transformed to the distribution of $L_{+,b}$ by transferring probability mass from higher values to lower values. This definition captures the notion that the outcome likelihood of positive individuals from group b is skewed lower than the outcome likelihood of positive individuals from a.

In a case study on FICO credit scores in Section 5, we find the minority group (blacks) is disadvantaged in features compared to the majority group (whites) (see Figure 2). There are many reasons that a group could be disadvantaged in features. Below, we go through a few potential causes.

Example 4.1 (Group membership explains away features)**.** Even if two groups are equally likely to have positive individuals, i.e., $\mathbb{P}(Y = 1 \mid G = a) = \mathbb{P}(Y = 1 \mid G = b)$, group b can still be disadvantaged compared to group a. Consider the graph below. Although the label Y is independent of the group G, the label Y is not independent of the group G once conditioned on the features X because the group G can provide an alternative reason for the observed features.



Concretely, let groups a and b be native and non-native speakers of english, X be the number of grammatical errors on an individual's job application, and Y be whether the individual is a qualified candidate. Negative individuals ($Y = 0$) are less meticulous when filling out their application and more likely to have grammatical errors. However, for individuals from group b there is another explanation for having grammatical errors – being a non-native speaker. Thus, positive individuals from group b end up with lower outcome likelihoods than those from a, even though they may be equally qualified.

Example 4.2 (Predicting base rates)**.** Suppose the rate of positives in group b is lower than that of group a: $\mathbb{P}(Y = 1 \mid G = b) < \mathbb{P}(Y = 1 \mid G = a)$. If there is a feature in the dataset that can be used as a proxy for predicting the group, such as zip code or name for predicting race, then the outcome likelihoods of positive individuals from group b can end up lower than those of positive individuals from group a because the features are simply predicting the base rate of each group.

Social gap increases. We now state and prove the main result showing that the social gap increases as the institution increases its threshold for acceptance. Before turning to the result, we introduce one technical requirement. The *likelihood condition* is that $\frac{\partial c_L(l,\tau)}{\partial l}$ is monotonically non-increasing in τ for $l, \tau \in [0, 1]$. When the cost function c is outcome monotonic, the likelihood condition is satisfied for a broad class of differentiable likelihood cost functions c_L, such as the following examples.

- Differentiable separable cost functions of the form $c_L(l, l') = \max(c_2(l') - c_1(l), 0)$ for $c_1, c_2 : [0, 1] \to \mathbb{R}_{\geq 0}$.
- Differentiable shift-invariant cost functions of the form

$$c_L(l, l') = \begin{cases} c_0(l' - l) & l < l' \\ 0 & l \geq l' \end{cases},$$

for $c_0 : [0, 1] \to \mathbb{R}_{\geq 0}$.

Notably, any linear cost $c_L(l, l') = \max(\alpha(l' - l), 0)$ where $\alpha > 0$ satisfies the likelihood condition.

Under the likelihood condition, we now show that the social gap increases as the institution increases its threshold for acceptance.

Theorem 4.1. *Let $\tau \in (0, 1]$ be the threshold of the classifier. If group b is disadvantaged in features compared to group a, and $\frac{\partial c_L(l,\tau)}{\partial l}$ is monotonically non-increasing in τ, then $\mathcal{G}(\tau)$ is positive and monotonically increasing over τ.*

Proof. By Lemma 2.1, any outcome monotonic cost function can be written as a cost over outcome likelihoods. Therefore, the social burden can be written as

$$\begin{aligned}\mathcal{B}_{+,g}(\tau) &= \mathbb{E}\left[\min_{f_\tau(x')=1} c(x, x') \mid Y = 1, G = g\right] \\ &= \int_0^\tau c_L(l, \tau)\, \mathrm{d}F_{g,1}(l),\end{aligned}$$

where $F_{+,g}$ denotes the CDF of the outcome likelihood $L_{+,g}$. Integrating by parts, we obtain a simple expression for $\mathcal{B}_{+,g}(\tau)$:

$$\begin{aligned}\mathcal{B}_{+,g}(\tau) &= \int_0^\tau c_L(l, \tau)\, \mathrm{d}F_{g,1}(l) \\ &= [c_L(l, \tau) F_{g,1}(l)]_0^\tau - \int_0^\tau \frac{\partial c_L(l, \tau)}{\partial l} F_{+,g}(l)\, \mathrm{d}l \\ &= -\int_0^\tau \frac{\partial c_L(l, \tau)}{\partial l} F_{+,g}(l)\, \mathrm{d}l\,,\end{aligned}$$

where the last line follows because $c_L(\tau, \tau) = 0$ and $F_{+,g}(0) = 0$. This expression for $\mathcal{B}_{+,g}(\tau)$ allows us to conveniently write the social gap as

$$\mathcal{G}(\tau) = \mathcal{B}_{+,b}(\tau) - \mathcal{B}_{+,a}(\tau) = \int_0^\tau \frac{\partial c_L(l, \tau)}{\partial l} (F_{+,a}(l) - F_{+,b}(l))\, \mathrm{d}l\,.$$

It is easy to observe $\mathcal{G}(\tau)$ is positive. By the monotonicity assumptions, $\frac{\partial c_L(l,\tau)}{\partial l} < 0$ for $l \in (0, \tau)$. Since group b is disadvantaged in features, $F_{+,a}(l) - F_{+,b}(l) < 0$ for $l \in (0, 1)$. Therefore, the social gap $\mathcal{G}(\tau) > 0$.

Now, we show $\mathcal{G}(\tau)$ is increasing in τ. Let $0 \le \tau < \tau' \le 1$. Then,the difference in the social gap is given by

$$\begin{aligned}\mathcal{G}(\tau') - \mathcal{G}(\tau) &= \int_0^\tau \frac{\partial\,(c_L(l, \tau') - c_L(l, \tau))}{\partial l} (F_{+,a}(l) - F_{+,b}(l))\, \mathrm{d}l \\ &+ \int_\tau^{\tau'} \frac{\partial c_L(l, \tau')}{\partial l} (F_{+,a}(l) - F_{+,b}(l))\, \mathrm{d}l.\end{aligned}$$

Since group b is disadvantaged in features, $(F_{+,a}(l) - F_{+,b}(l)) < 0$ for all l. By assumption, $\frac{\partial c_L(l,\tau)}{\partial l}$ is monotonically non-increasing in τ, so the first term is non-negative. Similarly, $\frac{\partial c_L(l,\tau')}{\partial l} < 0$ by monotonicity, so the second term is positive. Hence, $\mathcal{G}(\tau') - \mathcal{G}(\tau) > 0$, which establishes $\mathcal{G}(\tau)$ is monotonically increasing in τ. □

As a corollary, if the institution improves its utility beyond the non-strategic optimal classifier, then it also causes the social gap to increase.

Corollary 4.1. *Suppose group b is disadvantaged in features compared to group a, and $\frac{\partial c_L(l,\tau)}{\partial l}$ is monotonically non-decreasing in τ. Let $\tau \in (0,1]$ be a threshold and $\tau_0 = 0.5$ be the optimal non-strategic threshold. If $\mathcal{U}_\Delta(\tau) > \mathcal{U}_\Delta(\tau_0)$, then $\mathcal{G}(\tau) > \mathcal{G}(\tau_0)$.*

Proof. By Theorem 3.1, if $\mathcal{U}_\Delta(\tau) > \mathcal{U}_\Delta(\tau_0)$, then $\tau > \tau_0$. By Theorem 4.1, if $\tau > \tau_0$, then $\mathcal{G}(\tau) > \mathcal{G}(\tau_0)$. □

4.2 Different Costs

In Section 4.1, we showed that when two subpopulations have different feature distributions, the social burden can disproportionately fall on one group. In this section, we give show, even if the feature distributions of the two groups are exactly identical, the social burden can still disproportionately impact one group.

We have thus far assumed the existence of a cost function c that is uniform across groups a and b. For a variety of structural reasons, it is unlikely this assumption holds in practice. Rather, it is often the case that *different groups experience different costs* for changing their features.

When the cost for group b is systematically higher than the cost for group a, we prove group b incurs higher social burden than group a. Furthermore, if the institution improves its utility by increasing its threshold τ, then as a side effect it also increases the social gap between group b and a (Theorem 4.2).

Much of the prior work on fairness in classification focuses on preventing unfairness that can arise when different subpopulations have different distributions over features and labels [13, 15, 8]. Our result provides a reason to be concerned about the unfair impacts of a classifier even when two groups have identical initial distributions. Namely, that *it can be easier for one group to game the classifier than another.*

Formally, we say that group b is disadvantaged in cost compared to group a if the following condition holds.

Definition 4.3 (Disadvantaged in cost)**.** Let $c_g(x, x')$ be the cost for an individual from group g to adapt their features from x to x'. Group b is disadvantaged in cost if $c_b(x, x') \geq c_a(x, x')$ for all $x, x' \in X$.

Next, we give a variety of example scenarios of when a group can be disadvantaged in cost.

Example 4.3 (Opportunity Costs)**.** Many universities have adopted gender-neutral policies that stop the "tenure-clock" for a year for family-related reasons, e.g. childbirth. Ostensibly, no research is expected while the clock is stopped. The policies were made gender-neutral in an attempt to decrease the stigma women felt around taking time off for family reasons. However, the adoption of gender-neutral clocks actually *increased* the gap between the percentage of men and women who received tenure [2]. The suggested cause is that women still shoulder more of the burden of bearing and caring for children, compared to men. Men who stop their tenure clock are more productive during the period than women, who have a higher opportunity cost to doing research while raising a child.

Example 4.4 (Information Asymmetry)**.** A large portion of high-achieving, low-income students do not apply to selective colleges, despite the fact that these colleges are typically *less* expensive for them because of the financial aid they would receive [16]. This phenomenon seems to be due to low-income students having less access to information about college [17]. Since low-income students have more barriers to gaining information about college, it is natural to assume that, compared to their wealthier peers, they have a higher cost to strategically manipulating their admission features.

Example 4.5 (Economic Differences)**.** Consider a social media company that wishes to classify individuals as "influencers," either to more widely disseminate their content or to identify promising accounts for online marketing campaigns. Wealthy individuals can purchase followers or likes, whereas other groups have to increase these numbers organically [7]. Consequently, the costs to increasing one's popularity metric differs based on access to capital.

Finally, our main technical result shows that even when the distributions of groups a and b are identical, if group b is disadvantaged in cost, then when the institution increases its threshold for acceptance, it also increases the social gap between the two groups.

Theorem 4.2. *Suppose positive individuals from groups a and b have the same distribution over features, i.e, if $Z = (X \mid Y = 1)$, then Z is independent of the group G. If group b is disadvantaged in cost compared to group a, then the social gap $\mathcal{G}(\tau)$ is non-negative and monotonically non-decreasing in the threshold τ.*

Proof. The social burden to a group g can be written as

$$\mathcal{B}_{+,g}(\tau) = \int_{\mathcal{X}} \min_{x' : f_\tau(x')=1} c_g(x, x') p(X = x \mid Y = 1) \,\mathrm{d}x$$

because $X \mid Y = 1$ is independent of G. The social gap can then be expressed as

$$\begin{aligned}\mathcal{G}(\tau) &= \mathcal{B}_{+,b}(\tau) - \mathcal{B}_{+,a}(\tau) \\ &= \int_{\mathcal{X}} \left(\min_{x' : f_\tau(x')=1} c_a(x, x') - c_b(x, x') \right) p(X = x \mid Y = 1) dx \,.\end{aligned}$$

The gap in individual cost $\min_{x' : f_\tau(x')=1} c_a(x, x') - c_b(x, x')$ is always non-negative and is monotonically non-decreasing in τ, thus $\mathcal{G}(\tau)$ is non-negative and monotonically non-decreasing. □

5 Case Study: FICO Credit Data

We illustrate the impact of strategic classification on different subpopulations in the context of credit scoring and lending. FICO scores are widely used in the United States to predict credit worthiness. The scores themselves are derived from a proprietary classifier that uses features such as the number of open bank accounts that are susceptible to gaming and strategic manipulation.

We use a sample of 301,536 FICO scores derived from TransUnion TransRisk scores [24] and preprocessed by [15]. The scores X are normalized to lie between 0 and 100. An individual's outcome

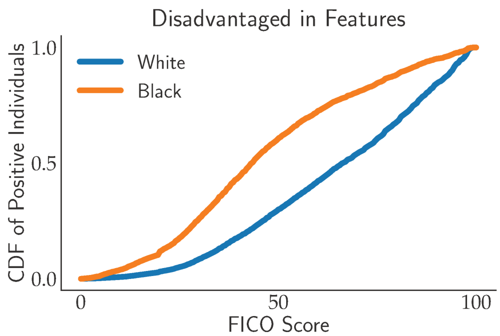


Figure 2: Comparison of the distribution of FICO scores among black and white borrowers who repaid their loans. Credit-worthy black individuals tend to have lower credit scores than credit-worthy white individuals. The comparison of the corresponding CDFs demonstrates our "disadvantaged in features" assumption holds.

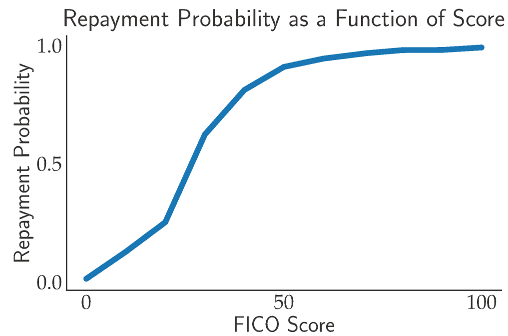


Figure 3: Repayment probability as a function of credit score. Crucially, the probability of repayment $\mathbb{P}(Y = 1 \mid x)$ is monotonically increasing in x.

is labeled as a *default* if she failed to pay a debt for at least 90 days on at least one account in the ensuing 18-24 month period. Default events are labeled with $Y = 0$, and otherwise repayment is denoted with $Y = 1$. The two subpopulations are given by race: $a = \texttt{white}$ and $b = \texttt{black}$.

We assume the credit lending institution accepts individuals based on a threshold on the FICO score. Using the normalized scale, a threshold of $\tau = 58$ is typically used to determine eligibility for prime rate loans [15]. Our results thus far have used thresholds on the outcome likelihood, rather than a score. However, as shown in Figure 3, the outcome likelihood is monotonic in the FICO score. Therefore, all our conditions and results can be validated using the score instead of the outcome likelihood.

5.1 Different Feature Distributions

In Section 4.1, we studied the scenario where the distribution of outcome likelihoods $\ell(X) = \mathbb{P}(Y = 1 \mid X)$ differed across subpopulations. In particular, if the likelihoods of the positive individuals in group B tend to be lower than the positive individuals in group A, then increasing strategic robustness increases the social gap between A and B.

Interestingly, such a skew in score distributions exists in the FICO data. Black borrowers who repay their loans tend to have lower FICO scores than white borrowers who repay their loans. In terms of the corresponding score CDFs, for every score x, $F_{+,\texttt{black}}(x) \geq F_{+,\texttt{white}}(x)$. Figure 2 demonstrates this observation.

When the score distribution among positive individuals is skewed, Theorem 4.1 guarantees the social gap between groups is increasing in the threshold under a reasonable cost model. Operationally, raising the loan threshold to protect against strategic behavior increases the relative burden on the black subgroup. To demonstrate this empirically, we use a coarse linear cost model, $c(x, x') = \max(\alpha(x' - x), 0)$ for some $\alpha > 0$. Since the probability of repayment $\mathbb{P}(Y = 1 \mid x)$ is monotonically

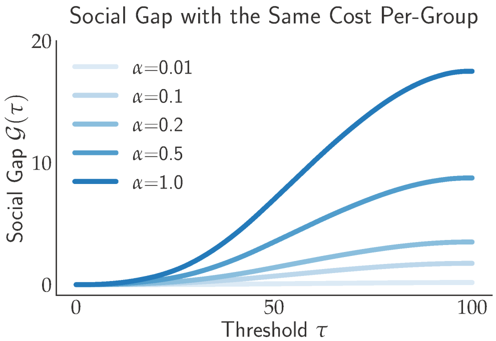


Figure 4: Impact of increasing the threshold τ on white and black credit applicants. When the cost to changing one's score α is small, increases to the threshold have only a small effect on the social gap. However, as α becomes large, even small increases to the threshold can create large discrepancies in social burden between the two groups.

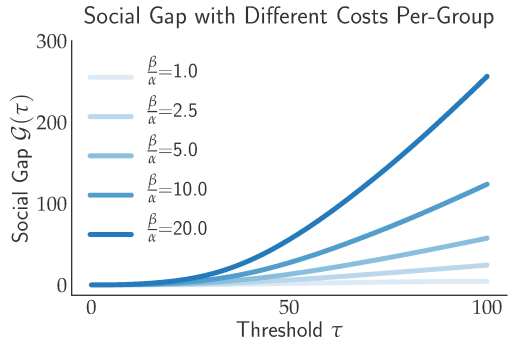


Figure 5: Impact of increasing the threshold τ on white and black credit applicants, under the assumption that both groups incur different costs for increasing their credit score. As the ratio of costs $\frac{\beta}{\alpha}$ increases, the social cost gap grows rapidly between the two groups.

increasing in x, the linear cost c satisfies the requisite outcome monotonicity conditions.

In Figure 4, we compute $\mathcal{G}(\tau)$ as τ varies from 0 to 100 for a range of different value of α. For any α, the social utility gap is increasing in τ. Moreover, as α (the cost of raising one's credit score) becomes large, the rate of increase in the social gap grows large as well.

5.2 Different Cost Functions

In Section 4.2, we demonstrated when two subpopulations are identically distributed, but incur different costs for changing their features, there is a non-trivial social gap between the two. In the context of the FICO scores, it's plausible blacks are both disadvantaged in features *and* experience higher costs for changing their scores. For instance, outstanding debt is an important component of FICO scores. One way to reduce debt is to increase earnings. However, a persistent black-white wage gap between the two subpopulations suggest increasing earnings is easier for group a than group b [11]. This setting is not strictly captured by our existing results, and we should expect the effects of both different costs functions and different feature distributions to compound and exacerbate the unfair impacts of strategic classification.

To illustrate this phenomenon, we again use a coarse linear cost model for both groups. Suppose group A has cost $c_A(x, x') = \max\{\alpha(x' - x), 0\}$ for some $\alpha > 0$ and group B has cost $c_B(x, x') = \max\{\beta(x' - x), 0\}$ for some $\beta \geq \alpha$. Since we are interested in the relative cost for each group, the key parameter controlling the rate of increase in $\mathcal{G}(\tau)$ is the ratio $\frac{\beta}{\alpha}$. In Figure 5, we show the social gap $\mathcal{G}(\tau)$ for various settings of $\frac{\beta}{\alpha}$. The social gap is always increasing as a function of τ, and the rate of increase grows large for even moderate values of $\frac{\beta}{\alpha}$. When $\frac{\beta}{\alpha}$ is large, even small increases in τ can disproportionately increase the social burden for the disadvantaged subpopulation.

6 Related Work

Strategic Classification Prior work on strategic classification focuses solely on the institution, primarily aiming to create high-utility solutions for the institution. Our work, on the other hand, studies the tradeoff between the institution's utility and the burden to the individuals being classified.

[14, 12, 4] give algorithms to compute the Stackelberg equilibrium, which corresponds to the extreme τ^* solution in our trade-off curves. Although the Stackelberg equilibrium leads to maximal institutional utility, we show that it also causes high social burden. We give several examples of when the high social burden induced by the Stackelberg equilibrium makes it an undesirable solution for the institution.

Rather than the Stackelberg equilibrium, others have also considered finding Nash equilibria of the game [5, 10]. [5] argue that since in practice people cannot optimally respond to the classifier, the Stackelberg solution tends to be too conservative, and thus a Nash equilibrium strategy is preferable. Our work provides a complementary reason to prefer Nash equilibria over the Stackelberg solution. Namely that (for a broad class of cost functions), any Nash equilibrium that is not equal to the Stackelberg equilibrium places lower social burden on individuals.

Finally, we focus on the setting where individuals are merely "gaming" their features, i.e., they do not improve their true label by adapting their features. However, if the classifier is able to incentivize strategic behavior that helps improve negative individuals, then the social burden placed on positive individuals may be considered acceptable. [20] studies how to design classifiers that produce such incentives.

Fairness Our work studies how *strategic classification* results in differing impacts to different subpopulations and is complementary to the large body of work studying the differing impacts of *classification* [23, 3].

The prior work on classification is primarily concerned with preventing unfairness that can arise due to subpopulations having differing distributions over features or labels [15, 13, 8]. We show that in the strategic setting, a classifier can have differing impact due to the subpopulations having differing distributions *or* differing costs to adapting their features. Therefore, when individuals are strategic, our work provides an additional reason to be concerned about the fairness of a classifier. Namely, that it can be easier for one group to game the classifier than another.

Furthermore, we show that if the institution modifies the classifier it uses to be more robust to strategic behavior, then it also as a side effect, increases the gap between the cost incurred by a disadvantaged subpopulation and an advantaged population. Thus, *strategic classification can exacerbate unfairness in classification.*

Our work is also complementary to [22], who also analyze how the institution's utility trades-off with the impact to individuals. They study the trade-off in the non-strategic setting and measure the impact of a classifier using a dynamics model of how individuals are affected by the classification they receive. We study the tradeoff in the strategic setting and measure the impact of a classifier by the cost of the strategic behavior induced by the classifier.

References

[1] Princeton and stanford are latest universities to drop sat/act writing test, Jul 2018. URL https://www.insidehighered.com/admissions/article/2018/07/09/princeton-and-stanford-are-latest-universities-drop-satact-writing.

[2] Heather Antecol, Kelly Bedard, and Jenna Stearns. Equal but inequitable: Who benefits from gender-neutral tenure clock stopping policies? In *American Economic Review*, forthcoming.

[3] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104 (3):671, 2016.

[4] Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[5] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 2012.

[6] Anthony P Carnevale, Nicole Smith, and Artem Gulish. Women can't win: Despite making educational gains and pursuing high-wage majors, women still earn less than men. 2018.

[7] Max Chafkin. Confessions of an instagram influencer, Nov 2016. URL https://www.bloomberg.com/news/features/2016-11-30/confessions-of-an-instagram-influencer.

[8] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 2017.

[9] Danielle Keats Citron and Frank Pasquale. The scored society: Due process for automated predictions. *Washington Law Review*, 89(1):1, 2014.

[10] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2004.

[11] Mary Daly, Bart Hobijn, Joseph H Pedtke, et al. Disappointing facts about the black-white wage gap. *FRBSF Economic Letter*, 2017:26, 2017.

[12] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *Conference on Economics and Computation (EC)*, 2018.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, 2012.

[14] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Conference on Innovations in Theoretical Computer Science (ITCS)*, 2016.

[15] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[16] Caroline Hoxby and Christopher Avery. The missing" one-offs": The hidden supply of high-achieving, low-income students. *Brookings Papers on Economic Activity*, 2013.

[17] Caroline Hoxby, Sarah Turner, et al. Expanding college opportunities for high-achieving, low income students. *Stanford Institute for Economic Policy Research Discussion Paper*, 2013.

[18] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. Aidr: Artificial intelligence for disaster response. In *International Conference on World Wide Web (WWW)*, 2014.

[19] Sarah Jeong. Bad romance, July 2018. URL `https://www.theverge.com/2018/7/16/17566276/cockygate-amazon-kindle-unlimited-algorithm-self-published-romance-novel-cabal`.

[20] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *arXiv preprint arXiv:1807.05307*, 2018.

[21] Haim Levy. Stochastic dominance and expected utility: survey and analysis. *Management science*, 1992.

[22] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

[23] Executive Office of the President, Cecilia Munoz, Domestic Policy Council Director, Megan (US Chief Technology Officer Smith (Office of Science, Technology Policy)), DJ (Deputy Chief Technology Officer for Data Policy, Chief Data Scientist Patil (Office of Science, and Technology Policy)). *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President, 2016.

[24] US Federal Reserve. Report to the congress on credit scoring and its effects on the availability and affordability of credit. 2007.

[25] Marilyn Strathern. 'improving ratings': audit in the british university system. *European review*, 5(3):305–321, 1997.

A Proof of Lemma 3.2

Proof. The lemma follows by proving the following properties about the Nash equilibrium strategies for the institution.

- The Stackelberg threshold τ^* is a Nash equilibrium strategy.
- All Nash equilibrium strategies lie in the interval $[\tau_0, \tau^*]$.
- If τ_N is a Nash equilibrium strategy, then all $\tau \in [\tau_N, \tau^*]$ are also Nash equilibrium strategies.

Together, the three properties imply that the set of institution equilibrium strategies is $[\tau_N, \tau^*]$ for some $\tau_N \in [\tau_0, \tau^*]$.

Before proceeding, recall the outcome likelihood $\ell(x) = \mathbb{P}(Y = 1 \mid X = x)$. For any threshold τ, define the strategic outcome likelihood $\ell_{\Delta_\tau}(x') = \mathbb{P}(Y = 1 \mid \Delta_\tau(X) = x')$. Let $\Delta_\tau(x)$ be the best response of individual x to the threshold τ. Define the set of individuals accepted for threshold τ and response Δ_τ by $\mathcal{A}_{\Delta_\tau}(\tau) = \{x : f_\tau(\Delta_\tau(x)) = 1\}$.

For the pair (τ, Δ_τ) to be a Nash equilibrium, τ must be a best response to the individual's best response $\Delta_\tau(x)$. With knowledge of the individual's response, $x' = \Delta_\tau(x)$, the institution's best response is to play a threshold τ' so that $x' \in \mathcal{A}_{\Delta_\tau}(\tau')$ iff $\ell_{\Delta_\tau}(x') \geq 0.5$. Therefore, to show (τ, Δ_τ) is a Nash equilibrium, we must show $\tau = \tau'$, i.e. $x' \in \mathcal{A}_{\Delta_\tau}(\tau)$ iff $\ell_{\Delta_\tau}(x') \geq 0.5$.

To verify the condition $x' \in \mathcal{A}_{\Delta_\tau}(\tau)$ iff $\ell_{\Delta_\tau}(x') \geq 0.5$, there are three cases to consider.

1. If $\ell(x) > \tau$, then $x \in \mathcal{A}_{\Delta_\tau}(\tau)$, $\Delta_\tau(x) = x$, and $\ell_{\Delta_\tau}(x) = \ell(x)$. Therefore, it suffices to check $\ell(x) \geq 0.5$.
2. If $\ell(x) < \tau$ and $c_L(\ell(x), \tau) > 1$, then $x \notin \mathcal{A}_{\Delta_\tau}(\tau)$ and $\Delta_\tau(x) = x$. In this case, it suffices to check $\ell(x) < 0.5$.
3. If $\ell(x) < \tau$ and $c_L(\ell(x), \tau) \leq 1$, then $\Delta_\tau(x) \in \mathcal{A}_{\Delta_\tau}(\tau)$, but $x \neq \Delta_\tau(x)$, so we must directly verify $\mathbb{P}(Y = 1 \mid c_L(\ell(x), \tau) \leq 1, \ell(x) \leq \tau) \geq 0.5$.

We now proceed to the proof.

First, we show the Stackelberg equilibrium $(\tau^*, \Delta_{\tau^*})$ is a Nash equilibrium. The Stackelberg threshold τ^* is the largest τ^* such that $c_L(0.5, \tau^*) \leq 1$. If $\ell(x) > \tau^*$, by monotonicity, $\ell(x) \geq 0.5$. If $\ell(x) < \tau^*$ and $c_L(\ell(x), \tau^*) > 1$, then $\ell(x) < 0.5$ by definition of τ^*. Similarly, if $\ell(x) < \tau^*$ and $c_L(\ell(x), \tau^*) \leq 1$, then $\ell(x) \geq 0.5$, so trivially $\mathbb{P}(Y = 1 \mid c_L(\ell(x), \tau) \leq 1, \ell(x) \leq \tau) \geq 0.5$. Hence, $(\tau^*, \Delta_{\tau^*})$ is a Nash equilibrium.

Next, we show that all Nash strategies must lie in the interval $[\tau_0, \tau^*]$.

1. Suppose $\tau < \tau_0 = 0.5$. For all x such that $\ell(\Delta_\tau(x)) = \tau$, $\ell(x) < 0.5$. Therefore, $\mathbb{P}(Y = 1 \mid c_L(\ell(x), \tau) \leq 1, \ell(x) \leq \tau) < 0.5$, so τ cannot be a Nash equilibrium strategy for the institution.
2. Suppose $\tau > \tau^*$. By definition, τ^* is the largest τ such that $c_L(0.5, \tau) \leq 1$. Thus, if $\tau > \tau^*$, there exists x with $\ell(x) < \tau^*$ and $c_L(\ell(x), \tau) > 1$, but $\ell(x) \geq 0.5$. Hence, τ cannot be a Nash strategy.

Finally, we show that if τ_N is a Nash equilibrium strategy, then so if τ for any $\tau \in [\tau_N, \tau^*]$. We consider each of the three cases in turn.

1. Suppose $\ell(x) > \tau$. Then $\ell(x) > \tau > \tau_N \geq \tau_0 = 0.5$.
2. Suppose $\ell(x) < \tau$ and $c_L(\ell(x), \tau) > 1$. Since $\tau \leq \tau^*$, by monotonicity, $1 < c_L(\ell(x), \tau) \leq c_L(\ell(x), \tau^*)$. However, $c_L(0.5, \tau^*) \leq 1$, so it follows that $\ell(x) < 0.5$.
3. Suppose $\ell(x) < \tau$ and $c_L(\ell(x), \tau) \leq 1$. Since τ_N is a Nash strategy, $\mathbb{P}(Y = 1 \mid c_L(\ell(X), \tau_N) \leq 1, \ell(X) \leq \tau_N) \geq 0.5$, and this probability is increasing in τ since $\ell(x) > 0.5$ for $\ell(x) > \tau_N$. Therefore, since $\tau \geq \tau_N$, $\mathbb{P}(Y = 1 \mid c_L(\ell(X), \tau) \leq 1, \ell(X) \leq \tau) \geq 0.5$.

Since each of the three cases are satisfied, any $\tau \in [\tau_N, \tau^*]$ is a Nash strategy.

We have now demonstrated each of the three properties outlined at the beginning, and the lemma follows. □