The Disparate Effects of Strategic Manipulation

Lily Hu
Harvard University
Cambridge, MA
lilyhu@g.harvard.edu

Nicole Immorlica
Microsoft Research
Cambridge, MA
nicimm@microsoft.com

Jennifer Wortman Vaughan
Microsoft Research
New York, NY
jenn@microsoft.com

ABSTRACT

When consequential decisions are informed by algorithmic input, individuals may feel compelled to alter their behavior in order to gain a system's approval. Models of agent responsiveness, termed "strategic manipulation," analyze the interaction between a learner and agents in a world where all agents are equally able to manipulate their features in an attempt to "trick" a published classifier. In cases of real world classification, however, an agent's ability to adapt to an algorithm is not simply a function of her personal interest in receiving a positive classification, but is bound up in a complex web of social factors that affect her ability to pursue certain action responses. In this paper, we adapt models of strategic manipulation to capture dynamics that may arise in a setting of social inequality wherein candidate groups face different costs to manipulation. We find that whenever one group's costs are higher than the other's, the learner's equilibrium strategy exhibits an inequality-reinforcing phenomenon wherein the learner erroneously admits some members of the advantaged group, while erroneously excluding some members of the disadvantaged group. We also consider the effects of interventions in which a learner subsidizes members of the disadvantaged group, lowering their costs in order to improve her own classification performance. Here we encounter a paradoxical result: there exist cases in which providing a subsidy improves only the learner's utility while actually making both candidate groups worse-off—even the group receiving the subsidy. Our results reveal the potentially adverse social ramifications of deploying tools that attempt to evaluate an individual's "quality" when agents' capacities to adaptively respond differ.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Theory of computation** → *Algorithmic game theory*;

KEYWORDS

fairness in machine learning; strategic classification

ACM Reference Format:
Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The Disparate Effects of Strategic Manipulation. In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3287560.3287597

FAT '19, January 29–31, 2019, Atlanta, GA, USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287597

1 INTRODUCTION

The expanding realm of algorithmic decision-making has not only altered the ways that institutions conduct their day-to-day operations, but has also had a profound impact on how individuals interface with these institutions. It has changed the ways we communicate with each other, receive crucial resources, and are granted important social and economic opportunities. Theoretically, algorithms have great potential to reform existing systems to become both more efficient and equitable, but as exposed by various high-profile investigations [2, 11, 23, 27], prediction-based models that make or assist with consequential decisions are, in practice, highly prone to reproducing past and current patterns of social inequality.

While few algorithmic systems are explicitly designed to be discriminatory, there are many underlying forces that drive such socially biased outcomes. For one, since most of the features used in these models are based on proxy, rather than causal, variables, outputs often reflect the various structural factors that bear on a person's life opportunities rather than the individualized characteristics that decision-makers often seek. Much of the previous work in algorithmic fairness has examined a particular undesirable proxy effect in which a classifier's features may be linked to socially significant and legally protected attributes like race and gender, interpreting correlations that have arisen due to centuries of accumulated disadvantage as genuine attributes of a particular category of people [13, 15, 19, 24].

But algorithmic models do not only generate outcomes that passively correlate with social advantages or disadvantages. These tools also provoke a certain type of reactivity, in which agents see a classifier as a guide to action and actively change their behavior to accord with the algorithm's preferences. On this view, classifiers both *evaluate* and *animate* their subjects, transforming static data into strategic responses. Just as an algorithm's use of certain features differentially advantages some populations over others, the room for strategic response that is inherent in many automated systems also naturally favors social groups of privilege. Admissions procedures that heavily weight SAT scores motivate students who have the means to take advantage of test prep courses and even take the exam multiple times. Loan approval systems that rely on existing lines of credit as an indication of creditworthiness encourage those who can to apply for more credit in their name.

Thus an algorithm that scores applicants to determine how a resource should be allocated sets a standard for what an ideal candidate's features ought to be. A responsive subject would look to alter how she appears to a classifier in order to increase her likelihood of gaining the system's approval. But since reactivity typically requires informational and material resources that are not equally

accessible to all, even when an algorithm draws on features that seem to arise out of individual effort, these metrics can be skewed to favor those who are more readily able to alter their features.

In the machine learning literature, agent reactivity to a classifier is termed "strategic manipulation." Since previous work in strategic classification has typically depicted agent-classifier interactions as antagonistic, such actions are usually viewed as distortions that aim to undermine a learner's classifier [5, 14]. As shown in Hardt et al. [14], a learner who anticipates these responses can, under certain formulations of agent costs, adapt to protect against the misclassification errors that would have resulted from manipulation, recovering an accuracy level that is arbitrarily close to the theoretical maximum. These results are welcome news for a learner who correctly assesses agents' best-responses. Indeed in most strategic manipulation models, agents are depicted as equally able to pursue manipulation, allowing the learner who knows their costs to accurately preempt strategic responses. While there are occasions in which agents do largely face homogenous costs—an even playing field—in many other social use cases of machine learning tools, agents do not encounter the same costs of altering the attributes that are ultimately observed and assessed by the classifier. As such, in this paper we ask, *"What are the effects of strategic classification and manipulation in a world of social stratification?"*

As in previous work in strategic classification, we cast the problem as a Stackelberg game in which the learner moves first and publishes her classifier before candidates best-respond and manipulate their features [1, 5, 7, 14]. But in contrast with the models in Brückner & Scheffer [5] and Hardt et al. [14], we formalize the setting of a society comprised of social groups that not only may differ in terms of distributions over unmanipulated features and true labeling functions but also face different costs to manipulation. This extra set of differences brings to light questions that favor an analysis that focuses on the welfares of the candidates who must contend with these classifiers: Do classifiers formulated with strategic behavior in mind impose disparate burdens on different groups? If so, how can a learner mitigate these adverse effects? The altered gameplay and outcomes of strategic classification beg questions of fairness that are intertwined with those of optimality.

Though our model is quite general, we obtain technical results that reveal important social ramifications of using classification in systems marked by deep inequalities and a potential for manipulation. Our analysis shows that, under our model, even when the learner knows the costs faced by different groups, her equilibrium classifier will always act to reinforce existing inequalities by mistakenly excluding qualified candidates who are less able to manipulate their features while also mistakenly admitting those candidates for whom manipulation is less costly, perpetuating the relative advantage of the privileged group. We delve into the cost disparities that generate such inevitable classification errors.

Next, we consider the impact of providing subsidies to lighten the burden of manipulation for the disadvantaged group. We find that such an intervention can improve the learner's classification performance as well as mitigate the extent to which her errors are inequality-reinforcing. However, we show that there exist cases in which providing subsidies enforces an equilibrium learner strategy that actually makes some individual candidates worse-off without making any better-off. Paradoxically, in these cases, paying a subsidy to the disadvantaged group actually benefits only the learner while both candidate groups experience a welfare decline! Further analysis of these scenarios reveals that, in many cases, all parties would have preferred a world in which manipulation of features was not possible for any candidates.

Our paper's agent-centric analysis views data points as representing individuals and classifications as impacting those individuals' welfares. This orientation departs from the dominant perspective in learning theory, which privileges a vendor's predictive accuracy, and instead evaluates classification regimes in light of the social consequences of the outcomes they issue. By incorporating insights and techniques from game theory and economics, domains that consider deeply the effects of various policies on agents' behaviors and outcomes, we hope to broaden the perspective that machine learning takes on socially-oriented tools. Presenting more democratically-inclined analysis has been central to the field of algorithmic fairness, and we hope our work sheds new light on this generic setting of classification with strategic agents.

1.1 Related Work

While many earlier approaches to strategic classification in the machine learning literature have tended to view learner-agent interactions as adversarial [3, 16], our work does not assume inherently antagonistic relationships, and instead, shares the Stackelberg game-theoretic perspective akin to that presented in Brückner & Scheffer [5] and built upon by Hardt et al. [14]. Departing from these models' focus on static prediction and homogeneous manipulation costs, Dong et al. [8] propose an online setting of strategic classification in which agents appear sequentially and have individual costs for manipulation that are unknown to the learner. Unlike our work, they take a traditional learner-centric view, whereas our concerns are with the welfare of the candidates.

Agent features and potential manipulations in the face of a learner classifier can also be interpreted as serving *informational* purposes. In the economics literature on signaling theory, agents interact with a principal—the counterpart to our learner—via signals that convey important information relevant to a particular task at hand. Classic works, such as Spence's paper on job-market signaling, focus their analysis on the varying quality of information that signals provide at equilibrium [25]. The emphasis in our analysis on different group costs shares features with a recent update to the signaling literature by Frankel & Kartik [12], who also distinguish between natural actions, corresponding to unmanipulated features in our model, and "gaming" ability, which operate similarly to our cost functions. The connection between gaming capacity and social advantage is also explicitly discussed in work by Esteban & Ray [10] who consider the effects of wealth and lobbying on governmental resource allocation. While most works in the economics signaling literature center on the decay of the informativeness of signals as gaming and natural actions become indistinguishable, some recent work in computer science has also considered the effect of costly signaling on mechanism design [17, 18]. In contrast to both of these perspectives, our work highlights the effect of manipulation on a learner's action and as a consequence, on the agents' welfares.

In independent, concurrent work appearing at the same conference, Milli et al. [22] also consider the social impacts of strategic classification. Whereas our model highlights the interplay between a learner's Stackelberg equilibrium classifier and agents' best-response manipulations at the feature level, their work traces the relationship between the learner's utility and the social burden, a measure of agents' manipulation costs. They show that an institution must select a point on the outcome curve that trades off its predictive accuracy with the social burden it imposes. In their model, an agent with an unmanipulated feature vector $\mathbf{x}$ has a likelihood $\ell(\mathbf{x})$ of having a positive label and can manipulate to any vector $\mathbf{y}$ with $\ell(\mathbf{y}) \leq \ell(\mathbf{x})$ at zero cost, or to $\mathbf{y}$ with $\ell(\mathbf{y}) > \ell(\mathbf{x})$ for a positive cost. This assumption, called "outcome monotonicity," allows them to reason about manipulations in (one-dimensional) likelihood space rather than feature space, since the optimal learner strategies amount to thresholds on likelihoods. In contrast, we allow features to be differently manipulable (perhaps a student can boost her SAT score via test prep courses, but can do nothing to change her grades from the previous year, and cannot freely obtain a higher SAT score in exchange for a worse record of extracurricular activities), which affects the forms of both the learner's equilibrium classifier and agents' best-response manipulations. Despite these differences in model and focus, their analysis yields results that are qualitatively similar to ours. Highlighting the differential impact of classifiers on social groups, they also find that overcoming stringent thresholds is more burdensome on the disadvantaged group.

2 MODEL FORMALIZATION

As in Brückner & Scheffer [5] and Hardt et al. [14], we formalize the Strategic Classification Game as a Stackelberg competition in which the learner moves first by committing to and publishing a binary classifier f. Candidates, who are endowed with "innate" features, best respond by manipulating their feature inputs into the classifier. Formally, a candidate is defined by her d-dimensional feature vector $\mathbf{x} \in X = [0,1]^d$ and group membership A or B, with A signifying the advantaged group and B the disadvantaged. Group membership bears on manipulation costs such that a candidate from group m who wishes to move from a feature vector $\mathbf{x}$ to a feature vector $\mathbf{y}$ must pay a cost of $c_m(\mathbf{y}) - c_m(\mathbf{x})$. We note that these cost function forms are similar to the class of separable cost functions considered in Hardt et al. [14]. We assume that higher feature values indicate higher quality to the learner, and thus restrict our attention to manipulations such that $\mathbf{y} \geq \mathbf{x}$, where the symbol $\geq$ signifies a component-wise comparison such that $\mathbf{y} \geq \mathbf{x}$ if and only if $\forall i \in [d]$, $y_i \geq x_i$. Throughout this paper, we study non-negative monotone cost functions such that the cost of manipulating from a feature vector $\mathbf{x}$ to a feature vector $\mathbf{y}$ increases as $\mathbf{x}$ and $\mathbf{y}$ get further apart.

To motivate this distinction between features and costs, consider the use of SAT scores as a signal of academic preparedness in the U.S. college admissions process. The high-stakes nature of the SAT has encouraged the growth of a test prep industry dedicated to helping students perform better on the exam. Test preparation books and courses, while also exposing students to content knowledge and skills that are covered on the SAT, promise to "hack" the exam by training students to internalize test-taking strategies based on the format, structure, and style of its questions. One can view SAT scores as a feature used by a learner building a classifier to select candidates with sufficient academic success according to some chosen standard. The existence of test prep resources then presents an opportunity for some applicants to inflate their scores, which might "trick" the tool into classifying the candidates as more highly qualified than they are in actuality. In this example, a candidate's strategic manipulation move refers to her investment in these resources, which despite improving her exam score, do not confer any genuine benefits to her level of academic preparation for college.

Just as access to test prep resources tends to fall along income and race lines, we view candidates' different abilities to manipulate as tied to their group membership. We model these group differences with respect to availability of resources and opportunity by enforcing a *cost condition* that orders the two groups. We suppose that for all $\mathbf{x} \in [0,1]^d$ and $\mathbf{y} \geq \mathbf{x}$,

$$c_A(\mathbf{y}) - c_A(\mathbf{x}) \leq c_B(\mathbf{y}) - c_B(\mathbf{x}). \tag{1}$$

Manipulating from a feature vector $\mathbf{x}$ to $\mathbf{y}$ is always at least as costly for a member of group B as it is for a member of group A. We believe our model's inclusion of this cost condition reflects an authentic aspect of our social world wherein one group is systematically disadvantaged with respect to a task in comparison to another.

In our setup, we also allow groups to have distinct probability distributions $\mathcal{D}_A$ and $\mathcal{D}_B$ over unmanipulated features and to be subject to different true labeling functions h_A and h_B defined as

$$h_A(\mathbf{x}) = \begin{cases} 1, & \forall \mathbf{x} \text{ such that } \sum_{i=1}^d w_{A,i} x_i \geq \tau_A, \\ 0, & \forall \mathbf{x} \text{ such that } \sum_{i=1}^d w_{A,i} x_i < \tau_A, \end{cases} \tag{2}$$

$$h_B(\mathbf{x}) = \begin{cases} 1, & \forall \mathbf{x} \text{ such that } \sum_{i=1}^d w_{B,i} x_i \geq \tau_B, \\ 0, & \forall \mathbf{x} \text{ such that } \sum_{i=1}^d w_{B,i} x_i < \tau_B. \end{cases} \tag{3}$$

We assume that $h_A(\mathbf{x}) = 1 \implies h_B(\mathbf{x}) = 1$ for all $\mathbf{x} \in [0,1]$. Returning to the SAT example, research has shown that scores are skewed by race even before factoring in additional considerations such as access to manipulation [6]. In such cases, the true threshold for the disadvantaged group is lower than that for the advantaged group. We leave this generality in our model to acknowledge and account for the influence that various social and historical factors have on candidates' unmanipulated features and not, we emphasize, as an endorsement of a view that groups are fundamentally different in ability. A formal description of the Strategic Classification Game with Groups is given in the following definition.

Definition 1 (Strategic Classification Game with Groups). *In the Strategic Classification Game with Groups, candidates with features $\mathbf{x} \in [0,1]^d$ and group memberships A or B are drawn from distributions $\mathcal{D}_A$ and $\mathcal{D}_B$. The population proportion of each group is given by p_A and p_B where $p_A + p_B = 1$. A candidate from group m pays cost $c_m(\mathbf{y}) - c_m(\mathbf{x})$ to move from her original features $\mathbf{x}$ to $\mathbf{y} \geq \mathbf{x}$. There exist true binary classifiers h_A and h_B, for candidates of each group. Probability distributions, cost functions, and true binary classifiers are all common knowledge. Gameplay proceeds in the following manner:*

1. *The learner issues a classifier f generating outcomes $\{0,1\}$.*
2. *Each candidate observes f and manipulates her features $\mathbf{x}$ to $\mathbf{y} \geq \mathbf{x}$.*

A group m candidate with features $\mathbf{x}$ who moves to $\mathbf{y}$ earns a payoff

$$f(\mathbf{y}) - (c_m(\mathbf{y}) - c_m(\mathbf{x})).$$

The learner incurs a penalty of

$$C_{FP} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m}[h_m(\mathbf{x}) = 0, f(\mathbf{y}) = 1] + C_{FN} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m}[h_m(\mathbf{x}) = 1, f(\mathbf{y}) = 0],$$

where C_{FP} and C_{FN} denote the cost of a false positive and a false negative respectively.

The learner looks to correctly classify candidates with respect to their original features $\mathbf{x}$, whereas each candidate hopes to manipulate her features to attain a positive classification, expending as little cost as possible in the process. Under this setup, candidates are only willing to manipulate their features if it flips their classification from 0 to 1 and if the cost of the manipulation is less than 1. We note that defining the utility of a positive classification to be 1 can be considered a scaling and thus is without loss of generality.

This learner-candidate interaction is very similar to that studied in Hardt et al. [14]. However, our inclusion of groups with distinct manipulation costs leads to an ambiguity regarding a candidate's initial features that does not exist when all candidates have an equal opportunity to manipulate. In very few cases can a vendor distinguish among candidates based on their group membership for the explicit purpose of issuing distinct classification policies, especially if that group category is a protected class attribute. As such, in our setup, we require that a learner publish a classifier that is not adaptive to different agents based on their group identities.

It is important to note that the positive results in Hardt et al.'s [14] formulation of the Strategic Classification Game, wherein for separable cost functions, the learner can attain a classification error at test-time that is arbitrarily close to the optimal payoff attainable, do not carry over into this setting of heterogeneous groups and costs. Even when $h_A = h_B$, the existence of different costs of agent manipulation, even when separable as in our model, introduces a base uncertainty to the learning problem that generates errors that cannot be extricated so long as the learner must publish a classifier that does not distinguish candidates based on their group memberships. Second, an analysis of the learner's strategy and performance, the perspective typically taken in most learning theory papers, contributes only a partial view of the total welfare effect of using classification in strategic settings. The main objective of this paper is to offer a more thorough and holistic inspection of all agents' outcomes, paying special heed to the different outcomes experienced by candidates of the two groups. Insofar as all social behaviors are impelled by goals, interests, and purposes, we should view data that is strategically generated to be the rule rather than the exception in social machine learning settings.

Remark on the assumption that h_A and h_B are known. Our assumption that the learner has knowledge of groups' true labeling functions is not central to our analysis. We make such an assumption to highlight the pure effect of groups' differential costs of manipulation on equilibrium gameplay and consequent welfares rather than the potential side effects due to a learner's noisy estimation of the true classifiers. Our general findings do not substantially rely on this feature of the model, and the overall results carry through into a setting in which the learner optimizes from samples.

Remark on unequal group costs. The differences in costs c_A and c_B encoded by the cost condition is not restricted to referring only to differences in the monetary cost of manipulation. Instead, as is common in information economics and especially signaling theory, "cost" reflects the multiplicity of factors that bear on the effort exertion required by feature manipulation [4, 21, 25, 26]. To demonstrate the generality of our formulation of distinct group costs, we show that the cost condition given in (1) is equivalent to a more explicit derivation of the choice that an agent faces when deciding whether to manipulate her feature.

A rational agent with feature $\mathbf{x}$ will only pursue manipulation if her value for a positive classification minus her cost of manipulation exceeds her value for a negative classification:

$$v(f(\mathbf{x}) = 0) \leq v(f(\mathbf{y}) = 1) - u(c(\mathbf{y}) - c(\mathbf{x})). \tag{4}$$

The monotone function u translates the costs borne by a candidate to manipulate from $\mathbf{x}$ to $\mathbf{y}$ into her "utility space," i.e., it reflects the value that she places on that expenditure. We can rewrite the previous inequality to be

$$c(\mathbf{y}) - c(\mathbf{x}) \leq u^{-1}\Big(v(f(\mathbf{y}) = 1) - v(f(\mathbf{x}) = 0)\Big). \tag{5}$$

Substituting in $k = u^{-1}\Big(v(f(\mathbf{y}) = 1) - v(f(\mathbf{x}) = 0)\Big)$, we have $c(\mathbf{y}) - c(\mathbf{x}) \leq k$. Since the same cost expenditure is valued more highly by the disadvantaged group than by the advantaged group, the function u is more convex for group B than for group A. Thus all else equal, we have $c_A(\mathbf{y}) - c_A(\mathbf{x}) \leq c_B(\mathbf{y}) - c_B(\mathbf{x})$ as desired. More generally, the functions v, c, and u may each be different for the groups. As such, the disadvantage encoded in the cost condition can arise due to differences in valuations of classifications (v), differences in costs (c), or differences in valuations of those costs (u).

3 EQUILIBRIUM ANALYSIS

We begin by studying agents' best-response strategies in the basic Strategic Manipulation Game with Groups in which candidates belong to one of two groups A and B, and the cost condition holds so that group B members face greater costs to manipulation than group A members. To build intuition, we first consider best-response strategies in the one-dimensional case in which candidates have features $x \in [0, 1]$ and group cost functions are of any non-negative monotone form. We then move on to consider the d-dimensional case in which candidate features are given as vectors $\mathbf{x} \in [0, 1]^d$ and manipulation costs are assumed to be linear.

3.1 One-dimensional Features

In the $d = 1$ case, the cost condition given in (1) may be written as $c'_A(x) \leq c'_B(x)$ for all $x \in [0, 1]$. Since the true decision boundaries are linear, in the one-dimensional case, they may be written as threshold functions where thresholds τ_A and τ_B are constants in $[0, 1]$ and for agents in group m, $h_m(x) = 1$ if and only if $x \geq \tau_m$. A university admissions decision based on a single score is an example of such a classifier. Although the SAT does not act as the sole determinant of admissions in the U.S., in countries such as Australia, Brazil, and China, a single exam score is often the only factor of applicant quality that is considered for admissions.

When the learner has access to τ_A and τ_B, and group costs c_A and c_B satisfy the cost condition, the following proposition characterizes

the space of undominated strategies for the learner who seeks to minimize any error-penalizing cost function.

Proposition 1 (One-D Undominated Learner Strategies). *Given group cost functions c_A and c_B and true label thresholds τ_A and τ_B where $\tau_B \leq \tau_A$, there exists a space of undominated learner threshold strategies $[\sigma_B, \sigma_A] \subset [0,1]$ where $\sigma_A = c_A^{-1}(c_A(\tau_A)+1)$ and $\sigma_B = c_B^{-1}(c_B(\tau_B)+1)$. That is, for any error penalties C_{FP} and C_{FN}, the learner's equilibrium classifier f is based on a threshold $\sigma \in [\sigma_B, \sigma_A]$ such that for all manipulated features y,*

$$f(y) = \begin{cases} 1, & \forall y \geq \sigma, \\ 0, & \forall y < \sigma. \end{cases} \tag{6}$$

To understand this result, first notice that if the learner were to face only those candidates from group A, she would achieve perfect classification by labeling as 1 only those candidates with unmanipulated feature $x \geq \tau_A$. This strategy is enacted by considering candidates' best-response manipulations. A rational candidate would only be willing to manipulate her feature if the gain she receives in her classification exceeds her costs of manipulation. The learner would like to guard against manipulations by candidates with $x < \tau_A$ but still admit candidates with $x \geq \tau_A$, so she considers the maximum manipulated feature y that is attainable by a rational candidate with $x = \tau_A$ who is willing to spend up to a cost of one in order to secure a better classification, as illustrated in Figure 1. The maximum such y value is σ_A, and thus, the learner sets a threshold at σ_A, admitting all those with $y \geq \sigma_A$ and rejecting all those with $y < \sigma_A$. The same reasoning applies to a learner facing only group B candidates, and the learner sets a threshold at σ_B, admitting all those candidates with $y \geq \sigma_B$ and rejecting all those with $y < \sigma_B$.

It can be shown that for all valid values of τ_A, τ_B, c_A, and c_B, necessarily $\sigma_B \leq \sigma_A$. Then all classifiers with threshold $\sigma < \sigma_B$ are dominated by σ_B, in the sense that for any arbitrary error penalties C_{FP} and C_{FN}, the learner would suffer higher costs by setting her threshold to be σ rather than σ_B. In the same way, all thresholds $\sigma > \sigma_A$ are dominated by σ_A, thus leaving $[\sigma_B, \sigma_A]$ to be the space of undominated thresholds. For an account of the full proof of this result (and all omitted proofs), see the appendix.

Even without committing to a particular learner cost function, the space of optimal strategies characterized in Proposition 1 leads to an important consequence. A rational learner in the Strategic Classification Game always selects a classifier that exhibits the following phenomenon: it mistakenly admits unqualified candidates from the group with lower costs and mistakenly excludes qualified candidates from the group with higher costs. This result is formalized in Proposition 2.

To state the proposition, the following definition is instructive. Whereas the true thresholds τ_A and τ_B are a function of unmanipulated features, the learner only faces candidate features that may have been manipulated. In order to make these observed features commensurable with τ_A and τ_B, it is helpful for the learner to "translate" a candidate's possibly manipulated feature y to its minimum corresponding original unmanipulated value.

Definition 2 (Correspondence with unmanipulated features). *For any observed candidate feature $y \in [0,1]$, the minimum corresponding unmanipulated feature is defined as*

$$\begin{aligned} \ell_A(y) &= \max\{0, c_A^{-1}(c_A(y)-1)\}, \\ \ell_B(y) &= \max\{0, c_B^{-1}(c_B(y)-1)\} \end{aligned} \tag{7}$$

for a candidate belonging to group A and group B respectively.

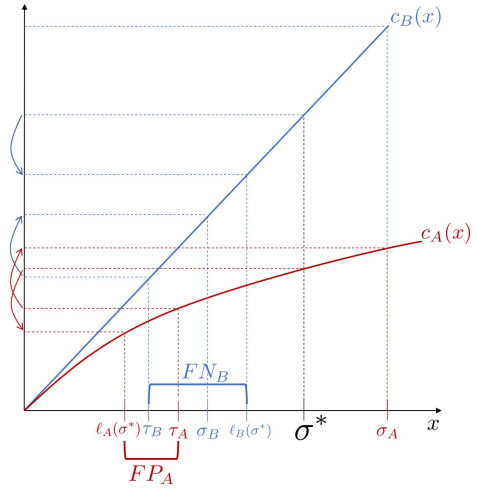

Figure 1: Group cost functions for a one-dimensional feature x. τ_A and τ_B signify true thresholds on unmanipulated features for group A and B, but a learner must issue a classifier on manipulated features. The threshold σ_A perfectly classifies group A candidates; σ_B perfectly classifies group B candidates. A learner selects an equilibrium threshold $\sigma^* \in [\sigma_B, \sigma_A]$, committing false positives on group A (red bracket) and false negatives on group B (blue bracket).

The corresponding values $\ell_A(y)$ and $\ell_B(y)$ are defined such that a candidate who presents feature y must have as her true unmanipulated feature $x \geq \ell_A(y)$ if she is a group A member and $x \geq \ell_B(y)$ if she is a group B member.

Proposition 2 (Learner's Cost in 1 Dimension). *A learner who employs a classifier f based on a threshold strategy $\sigma \in [\sigma_B, \sigma_A]$ only commits false positives errors on group A and false negatives errors on group B. The cost $C(\sigma)$ of such a classifier is*

$$C_{FN} p_B P_{x\sim\mathcal{D}_B}\Big[x \in [\tau_B, \ell_B(\sigma))\Big] + C_{FP} p_A P_{x\sim\mathcal{D}_A}\Big[x \in [\ell_A(\sigma), \tau_A)\Big],$$

where false negative errors entail penalty C_{FN}, and false positive errors entail penalty C_{FP}.

A learner who commits to classifying only one of the groups correctly bears costs given by the following corollaries.

Corollary 1. *A classifier based on σ_A perfectly classifies group A candidates and bears cost $C(\sigma_A) = C_{FN} p_B P_{x\sim\mathcal{D}_B}\Big[x \in [\tau_B, \ell_B(\sigma))\Big]$.*

Corollary 2. *A classifier based on σ_B perfectly classifies group B candidates and bears cost $C(\sigma_B) = C_{FP} p_A P_{x\sim\mathcal{D}_A}\Big[x \in [\ell_A(\sigma), \tau_A)\Big]$.*

Notice that the learner's errors always cut in the same direction—by unduly benefiting group A candidates and unduly rejecting group B candidates, these errors act to reinforce the existing social

inequality that had generated the unequal group cost conditions in the first place. Since these errors arise out of the asymmetric group costs of manipulation, the Strategic Classification Game can be viewed as an interactive model that itself perpetuates the relative advantage of group A over group B candidates.

Within the undominated region $[\sigma_B, \sigma_A]$, the equilibrium learner threshold σ^* is attained as the solution to the optimization problem

$$\sigma^* = \operatorname*{arg\,min}_{\sigma \in [\sigma_B, \sigma_A]} C(\sigma). \tag{8}$$

In the game's greatest generality where candidates are drawn from arbitrary probability distributions, groups bear any costs that abide by the cost condition, and the learner has arbitrary error penalties, one cannot specify the equilibrium learner threshold σ^* any further. However, under some special cases of candidate cost functions and probability distributions, the equilibrium threshold can be characterized more precisely. Specifically, when candidates from both groups are assumed to be drawn from a uniform distribution over unmanipulated features in $[0, 1]$, an error-minimizing learner seeks a threshold value σ^* that minimizes the length of the interval of errors, given by the following quantity:

$$\sigma^* = \operatorname*{arg\,min}_{\sigma \in [\sigma_B, \sigma_A]} \ell_B(\sigma) - \ell_A(\sigma).$$

From here, one natural assumption of candidate cost functions would have that groups A and B bear costs that are proportional to each other. In this case, the curvature of the cost functions is determinative of a learner's equilibrium threshold.

Proposition 3. *Suppose group cost functions are proportional such that $c_A(x) = qc_B(x)$ for $q \in (0, 1)$, that $\mathcal{D}_A$ and $\mathcal{D}_B$ are uniform on $[0, 1]$, and that $C_{FN} = C_{FP}$ and $p_A = p_B = \frac{1}{2}$. Let σ^* be the learner's equilibrium threshold.*
When cost functions are strictly concave, $\sigma^* = \sigma_B$. When cost functions are strictly convex, $\sigma^* = \sigma_A$. When cost functions are affine, the learner is indifferent between all $\sigma^* \in [\sigma_B, \sigma_A]$.

3.2 General d-Dimensional Feature Vectors

In the general d-dimensional case of the Strategic Classification Game, candidates are endowed with features that are given by a vector $\mathbf{x} \in [0, 1]^d$ and can choose to manipulate and present any feature $\mathbf{y} \geq \mathbf{x}$ to the learner. In this section, we consider optimal learner and candidate strategies when group costs are linear such that they may be written as

$$c_A(\mathbf{x}) = \sum_{i=1}^{d} c_{A,i} x_i; \quad c_B(\mathbf{x}) = \sum_{i=1}^{d} c_{B,i} x_i \tag{9}$$

for groups A and B respectively. Now, the cost condition $c_A(\mathbf{y}) - c_A(\mathbf{x}) \leq c_B(\mathbf{y}) - c_B(\mathbf{x})$ for all $\mathbf{y} \geq \mathbf{x}$—defined component-wise as before—implies that $\forall i \in [d]$, $c_{A,i} \leq c_{B,i}$. In d dimensions, the true classifiers h_A and h_B have linear decision boundaries such that for a group A candidate with feature $\mathbf{x}$,

$$h_A(\mathbf{x}) = \begin{cases} 1 & \sum_{i=1}^{d} w_{A,i} x_i \geq \tau_A, \\ 0 & \sum_{i=1}^{d} w_{A,i} x_i < \tau_A, \end{cases} \tag{10}$$

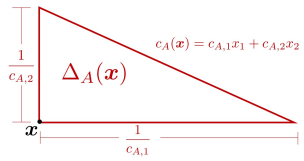

Figure 2: The forward simplex. A candidate in group A with unmanipulated feature vector $\mathbf{x}$ can manipulate to reach any feature vector $\mathbf{y} \in \Delta_A(\mathbf{x})$ at a cost of at most 1.

and for a group B candidate with feature $\mathbf{x}$,

$$h_B(\mathbf{x}) = \begin{cases} 1 & \sum_{i=1}^{d} w_{B,i} x_i \geq \tau_B, \\ 0 & \sum_{i=1}^{d} w_{B,i} x_i < \tau_B. \end{cases} \tag{11}$$

We assume that all components x_i contribute positively to an agent's likelihood of being classified as 1 so that $w_{A,i}, w_{B,i} \geq 0$ for all i. To ensure that the cost of manipulation is always non-negative, all cost coefficients are positive: $c_{B,i}, c_{A,i} \geq 0$ for all $i \in [d]$.

A candidate may now manipulate any combination of the d components of her initial feature $\mathbf{x}$ to reach the final feature $\mathbf{y}$ that she presents to the learner. Despite this increased flexibility on the part of the candidate, we are still able to characterize the performance of undominated learner classifiers, generalizing the result in Proposition 2. All potentially optimal classifiers exhibit the same inequality-reinforcing property inherent within the one-dimensional interval of undominated threshold strategies, trading off false positives on group A candidates with false negatives on group B candidates. Before we formally present this result, we first describe candidates' best-response strategies. Here, a geometric view of the space of potential manipulations is informative.

Suppose a candidate endowed with a feature vector $\mathbf{x}$ faces costs $\sum_{i=1}^{d} c_i x_i$ and is willing to expend a total cost of 1 for manipulation. Then she can move to any $\mathbf{y} \geq \mathbf{x}$ contained within the d-simplex with orthogonal corner at $\mathbf{x}$ and remaining vertices at $\mathbf{x} + \frac{1}{c_i}\mathbf{e}_i$ where $\mathbf{e}_i$ is the ith standard basis vector. This region is given by

$$\Delta(\mathbf{x}) = \left\{ \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i} \mathbf{e}_i \in [0, 1]^d \,\middle|\, \sum_{i=1}^{d} t_i \leq 1 \,;\, t_i \geq 0 \;\forall i \right\}. \tag{12}$$

$\Delta(\mathbf{x})$, depicted in Figure 2, gives the space of potential movement for a candidate with unmanipulated feature $\mathbf{x}$ who is willing to expend a total cost of 1. Notice that t_i can be interpreted as the cost that a candidate expends on movement in the ith direction. Thus $\sum_{i=1}^{d} t_i$ gives the total cost of manipulation. Moving beyond the range of possible moves, in order to describe how a rational candidate will best-respond to a learner, we must consider the published classifier.

Suppose a learner publishes a classifier f based on a hyperplane $\sum_{i=1}^{d} g_i y_i = g_0$, so that $f(\mathbf{y}) = 1$ if and only if $\sum_{i=1}^{d} g_i y_i \geq g_0$. A best-response manipulation occurs along the direction that generates the greatest increase in the value $\sum_{i=1}^{d} g_i(y_i - x_i)$ for the least cost. As such, a candidate will move in any directions $i \in \arg\max_{i \in [d]} \frac{g_i}{c_i}$. This result is formalized in the following lemma.

Lemma 1 (d-D Candidate Best Response). *Suppose a learner publishes the classifier $f(\mathbf{y}) = 1$ if and only if $\sum_{i=1}^{d} g_i y_i \geq g_0$. Consider a candidate with unmanipulated feature vector $\mathbf{x}$ and linear*

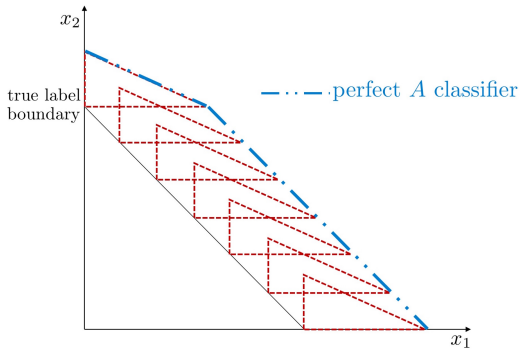

Figure 3: A perfect classifier for group A. Every candidate with unmanipulated feature vector $\mathbf{x}$ on or above the true decision boundary for group A is able to manipulate to a point $\mathbf{y} \in \Delta_A(\mathbf{x})$ on or above the blue decision boundary depicted here. No candidate with an unmanipulated feature vector below the true decision boundary is able to do so. The kink in the blue decision boundary arises due to the restriction of features to $[0,1]^d$. A perfect classifier for group A does not need to have this kink; for example, a more lenient perfect classifier can be formed by "straightening" it out.

costs $\sum_{i=1}^d c_i x_i$. *If* $f(\mathbf{x}) = 1$ *or if for all* $i \in [d]$, $f(\mathbf{x} + \frac{1}{c_i}\mathbf{e}_i) = 0$, *the candidate's best response is to set* $\mathbf{y} = \mathbf{x}$. *Otherwise, letting* $K = \arg\max_{i \in [d]} \frac{g_i}{c_i}$, *her manipulation takes the form*

$$y = \mathbf{x} + \sum_{i=1}^{d} \frac{t_i}{c_i}\mathbf{e}_i$$

for any $\mathbf{t}$ *such that* $t_i \geq 0$ *for all* $i \in [d]$, $t_i = 0$ *for all* $i \notin K$, *and* $\sum_{i=1}^d g_i(x_i + \frac{t_i}{c_i}) = g_0$.

While in the d-dimensional case, a candidate has many more choices of manipulation directions to pursue, a best response strategy will always lead her to increase her feature in those components that are most valued by the learner and least costly for manipulation. That is, she behaves according to a "bang for your buck" principle, in which the optimal manipulations are in the direction or directions where the ratio $\frac{g_i}{c_i}$ is highest.

Despite the fact that the optimal manipulation may not be unique, as in the cases where there are multiple equivalently good directions for a candidate to move in, a learner who knows candidates' costs can still anticipate best-response manipulations and avoid errors on that group. As such, we are once again able to construct a perfect classifier for candidates of group A and a perfect classifier for candidates of group B.

Theorem 1 (d-D Space of Dominant Learner Strategies). *In the general d-dimensional Strategic Classification Game with linear costs, there exists a classifier that perfectly classifies group A and a classifier that perfectly classifies group B. All undominated classifiers commit no false positive errors on group A and no false negative errors on group B.*

A full exposition of the proof appears in the appendix, but here we present an abbreviated explanation of the result.

For each group m, the learner computes an optimal boundary that perfectly classifies all of its members by considering the set of simplices $\{\Delta_m(\mathbf{x})\}$ anchored at the vectors $\bar{\mathbf{x}}$ that satisfy $\mathbf{w}_m^\intercal \bar{\mathbf{x}} = \tau_m$ and drawing the strictest hyperplane that intersects each simplex. That is for all hyperplanes $g_i : \sum_{j=1}^d g_{i,j}x_j = g_{i,0}$ that are constructed to intersect each simplex, then $g_1 : \sum_{j=1}^d g_{1,j}x_j = g_{1,0}$ is the strictest if for all $\mathbf{x} \in [0,1]^d$,

$$\sum_{j=1}^{d} g_{1,j}x_j = g_{1,0} \implies \sum_{j=1}^{d} g_{i,j}x_j = g_{i,0} \geq g_{j,0}$$

for all g_i. Due to the cost ordering, for any $\mathbf{x} \in [0,1]^d$, $\Delta_B(\mathbf{x}) \subseteq \Delta_A(\mathbf{x})$, and thus wherever a comparison is possible, the group A boundary is at least as strict as the group B boundary. Figure 3 gives a visualization of a boundary formed by connecting the simplices $\Delta(\bar{\mathbf{x}})$; the corresponding classifier perfectly classifies the group.

As in the one-dimensional general costs case, learner strategies necessarily entail inequality-reinforcing classifiers: a rational learner equipped with any error-penalizing cost function will select an equilibrium strategy that trades off undue optimism with respect to group A for undue pessimism with respect to group B. We note that except in the extreme case in which there exists a perfect classifier for all candidates in the population, this result implies that the classifier for group A issues false negatives on group B, and the classifier for group B issues false positives on group A. In order to formalize this result, we would like to generalize the idea behind the minimum correspondence unmanipulated features given by $\ell_A(\cdot)$ and $\ell_B(\cdot)$ in (7) for general d-dimensions and linear costs.

A learner who observes a possibly manipulated feature vector $\mathbf{y}$ must consider the space of unmanipulated feature vectors that the candidate could have had. Thus we can make use of the simplex idea of potential manipulation; however in this case, the learner seeks to project a simplex "backward" to "undo" the potential candidate manipulation. Since groups are subject to different costs, simplices $\Delta_A^{-1}(\mathbf{y})$ and $\Delta_B^{-1}(\mathbf{y})$—a depiction is given in Figure 4—which represent the region from where a candidate could have manipulated, will differ based on the candidate's group membership, with

$$\Delta_A^{-1}(\mathbf{y}) = \left\{\mathbf{y} - \sum_{i=1}^{d} \frac{t_i}{c_{A,i}}\mathbf{e}_i \in [0,1]^d \middle| \sum_{i=1}^{d} t_i \leq 1 \,;\, t_i \geq 0 \;\forall i\right\}, \quad (13)$$

$$\Delta_B^{-1}(\mathbf{y}) = \left\{\mathbf{y} - \sum_{i=1}^{d} \frac{t_i}{c_{B,i}}\mathbf{e}_i \in [0,1]^d \middle| \sum_{i=1}^{d} t_i \leq 1 \,;\, t_i \geq 0 \;\forall i\right\}. \quad (14)$$

We can now use these constructs in order to define d-dimensional generalizations of $\ell_A(\mathbf{y})$ and $\ell_B(\mathbf{y})$.

Definition 3 (Correspondence with Unmanipulated Features in d-D). *For any observed candidate feature* $\mathbf{y} \in [0,1]^d$, *the minimum corresponding unmanipulated feature vectors are given by*

$$\ell_A(\mathbf{y}) = \left\{\mathbf{x} \in \Delta_A^{-1}(\mathbf{y}) \cap [0,1]^d \middle| \nexists \hat{\mathbf{x}} \in \Delta_A^{-1}(\mathbf{y}) \text{ such that } \hat{\mathbf{x}} < \mathbf{x}\right\}, \quad (15)$$

$$\ell_B(\mathbf{y}) = \left\{\mathbf{x} \in \Delta_B^{-1}(\mathbf{y}) \cap [0,1]^d \middle| \nexists \hat{\mathbf{x}} \in \Delta_B^{-1}(\mathbf{y}) \text{ such that } \hat{\mathbf{x}} < \mathbf{x}\right\} \quad (16)$$

for a candidate belonging to group A and group B respectively.

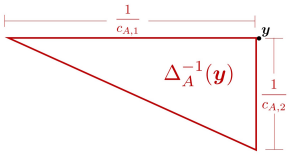

Figure 4: The backward simplex. A candidate in group A with manipulated feature vector $\mathbf{y}$ could have started with any feature vector $\mathbf{x} \in \Delta_A^{-1}(\mathbf{y})$ and paid a cost of at most 1.

The corresponding values $\ell_A(\mathbf{y})$ and $\ell_B(\mathbf{y})$ are defined such that a candidate who presents feature $\mathbf{y}$ must have had a true unmanipulated feature vector $\mathbf{x} \geq \bar{\mathbf{x}}$ for some $\bar{\mathbf{x}} \in \ell_A(\mathbf{y})$ if she is a group A member and $\mathbf{x} \geq \bar{\mathbf{x}}$ for some $\bar{\mathbf{x}} \in \ell_B(\mathbf{y})$ if she is a group B member.

For any hyperplane decision boundary g containing vectors $\mathbf{y}$, the minimum corresponding feature vectors given by $\ell_A(\mathbf{y})$ and $\ell_B(\mathbf{y})$ are helpful for determining the effective thresholds that g generates on unmanipulated features for groups A and B.

Lemma 2. *Suppose a learner classifier f is based on a hyperplane $g : \sum_{i=1}^d g_i x_i = g_0$. Construct the set*

$$\mathcal{L}_m(g) = \left\{ \arg\min_{\mathbf{x} \in \ell_m(\mathbf{y})} \sum_{i=1}^d g_i x_i \middle| \forall \mathbf{y} \text{ s. t. } \sum_{i=1}^d g_i y_i = g_0 \right\} \quad (17)$$

Then a group m agent with feature $\mathbf{x}$ can move to some $\mathbf{y}$ with $f(\mathbf{y}) = 1$ and $c_m(\mathbf{y}) - c_m(\mathbf{x}) \leq 1$ if and only if $\mathbf{x} \geq \ell$ for some $\ell \in \mathcal{L}_m(g)$.

By definition, for any two $\ell_1, \ell_2 \in \mathcal{L}_m(g)$,

$$\sum_{i=1} g_i \ell_{1,i} = \sum_{i=1} g_i \ell_{2,i} = g_0 - \frac{g_{k_m}}{c_{m,k_m}},$$

where $k_m \in \arg\max_{i=[d]} \frac{g_i}{c_{m,i}}$. Thus a learner who cares only about the true label of presented features, will construct her decision boundary g such that all $\ell \in \mathcal{L}_m(g)$ have the same true label.

A cost-minimizing learner who publishes a classifier f based on a hyperplane g on manipulated features will commit errors on those candidates with unmanipulated features $\mathbf{x} \in [0,1]^d$ contained within the boundaries given by $\mathcal{L}_A(g)$ and $\mathcal{L}_B(g)$. This space can be understood as the d-dimensional generalization of the $[\ell_A(\sigma), \ell_B(\sigma)]$ error interval in one-dimension.

Proposition 4 (Learner's Cost in d Dimensions). *A learner who publishes an undominated classifier f based on a hyperplane $\boldsymbol{g}^\intercal \mathbf{x} = g_0$ can only commit false positives on group A candidates and false negatives on group B candidates. The cost of such a classifier is*

$$C_{FN} P_{x \sim \mathcal{D}_B} \left[\mathbf{x} \in \left(\boldsymbol{g}^\intercal \mathbf{x} < g_0 - \frac{g_{k_B}}{c_{k_B}} \bigcap \mathbf{w}_B^\intercal \mathbf{x} \geq \tau_B \right) \right]$$
$$+ C_{FP} P_{x \sim \mathcal{D}_A} \left[\mathbf{x} \in \left(\mathbf{w}_A^\intercal \mathbf{x} < \tau_A \bigcap \boldsymbol{g}^\intercal \mathbf{x} \geq g_0 - \frac{g_{k_A}}{c_{k_A}} \right) \right],$$

where $k_B \in \arg\max_{i \in [d]} \frac{g_i}{c_{B,i}}$ and $k_A \in \arg\max_{i \in [d]} \frac{g_i}{c_{A,i}}$.

4 LEARNER SUBSIDY STRATEGIES

Since in our setting, the learner's classification errors are directly tied to unequal group costs, we ask whether she would be willing to subsidize group B candidates in order to shrink the manipulation gap between the two groups and as a result, reduce the number of errors she commits. In this section, we formalize subsidies as interventions that a learner can undertake to improve her classification performance. Although in many high-stakes classification settings, the barriers that make manipulation differentially accessible are non-monetary—such as time, information, and social access—in this section, we consider subsidies that are monetary in nature to alleviate the financial burdens of manipulation.

We introduce these subsidies for the purpose of analyzing their effects on not only the learner's classification performance but also candidate groups' outcomes. Since subsidies mitigate the inherent disparities in groups' costs and increase access to manipulation, one might expect that their implementation would surely improve group B's overall welfare. In this section, we show that in some cases, optimal subsidy interventions can surprisingly have the effect of lowering the welfare of candidates from *both* groups without improving the welfare of even a single candidate.

4.1 Subsidy Formalization

There are different ways in which a learner might choose to subsidize candidates costs. In the main text of this paper, we focus on subsidies that reduce each group B candidate's costs such that the agent need only pay a β fraction of her original manipulation cost.

Definition 4 (Proportional subsidy). *Under a proportional subsidy plan, the learner pays a proportion $1 - \beta$ of each group B candidate's cost of manipulation for some $\beta \in [0,1]$. As such, a group B candidate who manipulates from an initial feature vector $\mathbf{x}$ to a final feature vector $\mathbf{y}$ bears a cost of $\beta\big(c_B(\mathbf{y}) - c_B(\mathbf{x})\big)$.*

In the appendix, we also introduce flat subsidies in which the learner absorbs up to a flat α amount from each group B candidate's costs, leaving the candidate to pay $\max\{0, c_B(\mathbf{y}) - c_B(\mathbf{x}) - \alpha\}$. Similar results to those shown in this section hold for flat subsidies.

When considering proportional subsidies, the learner's strategy now consists of both a choice of β and a choice of classifier f to issue. The learner's goal is to minimize her penalty

$$C_{FP} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m} \left[h_m(\mathbf{x}) = 0, f(\mathbf{y}) = 1 \right]$$
$$+ C_{FN} \sum_{m \in \{A,B\}} p_m P_{\mathbf{x} \sim \mathcal{D}_m} \left[h_m(\mathbf{x}) = 1, f(\mathbf{y}) = 0 \right] + \lambda cost(f, \beta),$$

where $cost(f, \beta)$ is the monetary cost of the subsidy, C_{FP} and C_{FN} denote the cost of a false positive and a false negative respectively as before, and $\lambda \geq 0$ is some constant that determines the relative weight of misclassification errors and subsidy costs for the learner.

For ease of exposition, the remainder of the section is presented in terms of one-dimensional features. In Section A.3.1 of the appendix, we show that in many cases, the d-dimensional linear costs setting can be reduced to this one-dimensional setting.

As an analog of (7), we define $\ell_B^\beta(y) = (\beta c_B)^{-1}(\beta c_B(y) - 1)$, giving the minimum corresponding unmanipulated feature x for any observed feature y. Under the proportional subsidy, for a given y, the group B candidate must have $x \geq \ell_B^\beta(y)$. From this, we define σ_B^β such that $\ell_B^\beta(\sigma_B^\beta) = \tau_B$.

In order to compute the cost of a subsidy plan, we must determine the number of group B candidates who will take advantage of a

given subsidy benefit. Since manipulation brings no benefit in itself, candidates will only choose to manipulate and use the subsidy if it will lead to a positive classification. For a published classifier f with threshold σ, we then have

$$cost(f, \beta) = \Big(1 - \beta\Big) \int_{\ell_B^\beta(\sigma)}^{\sigma} \Big(c_B(\sigma) - c_B(x)\Big) P_{x \sim \mathcal{D}_B}(x) dx.$$

Although the learner's optimization problem can be solved analytically for various values of λ, we are primarily interested in taking a welfare-based perspective on the effects of various classification regimes on both the learner and candidate groups. In the following section, we analyze how the implementation of a subsidy plan can alter a learner's classification strategy and consider the potential impacts of such policies on candidate groups.

4.2 Group Welfare Under Subsidy Plans

While a learner would choose to adopt a subsidy strategy primarily in order to reduce her error rate, offering cost subsidies can also be seen as an intervention that might equalize opportunities in an environment that by default favors those who face lower costs. That is, if costs are keeping group B down, then one might believe that reducing costs will surely allow group B a fairer shot at manipulation, and, as a result, a fairer shot at positive classification. Alas we find that mitigating cost disparities by way of subsidies does not necessarily lead to better outcomes for group B candidates. In fact, an optimal subsidy plan can actually reduce the welfares of *both* groups. Paradoxically, in some cases, the subsidy plan boosts only the learner's utility, whereas every individual candidate from both groups would have preferred that she offer no subsidies at all.

The following theorem captures the surprising result that subsidies can be harmful to all candidates, even those from the group that would appear to benefit.

Theorem 2 (Subsidies can harm both groups). *There exist cost functions c_A and c_B satisfying the cost conditions, learner distributions $\mathcal{D}_A$ and $\mathcal{D}_B$, true classifiers with threshold τ_A and τ_B, population proportions p_A and p_B, and learner penalty parameters C_{FN}, C_{FP}, and λ, such that no candidate in either group has higher payoff at the equilibrium of the Strategic Classification Game with proportional subsidies compared with the equilibrium of the Strategic Classification Game with no subsidies, and some candidates from both group A and group B are strictly worse off.*

We note that a slightly weaker version of the theorem holds for flat subsidies. In particular, there exist cases in which some individual candidates have higher payoff at the equilibrium of the Strategic Classification Game with flat subsidies compared with the equilibrium with no subsidies, but both group A and group B candidates have lower payoffs on average with the subsidies.

To prove the theorem, it suffices to give a single case in which both candidate groups are harmed by the use of subsidies. However, to illustrate that this phenomenon does not arise only as a rare corner case, we provide one such example here plus two in the appendix, and discuss general conditions under which this occurs. In each example, we consider a particular instance of the Strategic Classification Game and compare the welfares of candidates at equilibrium when the learner is able to select a proportional subsidy with their welfares at equilibrium when no subsidy is allowed.

Example 1. *Suppose that a learner is error-minimizing such that $C_{FN} = C_{FP} = 1$ and $\lambda = \frac{3}{4}$. Suppose that unmanipulated features for both groups are uniformly distributed with $p_A = p_B = \frac{1}{2}$. Let group cost functions be given by $c_A(x) = 8\sqrt{x} + x$ and $c_B(x) = 12\sqrt{x}$; note that the cost condition $c'_A(x) < c'_B(x)$ holds for $x \in [0, 1]$. Let the true group thresholds be given by $\tau_A = 0.4$ and $\tau_B = 0.3$.*

When subsidies are not allowed, the learner chooses a classifier with threshold $\sigma^* = \sigma_B \approx 0.398$ at equilibrium. This threshold perfectly classifies all candidates from group B, while permitting false positives on candidates from group A with features $x \in [0.272, 0.4)$.

If the learner decides to implement a proportional subsidies plan, at equilibrium the learner chooses a classifier with threshold $\sigma^*_{prop} = \sigma_A \approx 0.546$ and a subsidy parameter $\beta^* = 0.558$. Her new threshold now correctly classifies all members of group A, while committing false negatives on group B members with features $x \in [0.3, 0.348)$.

Some candidates in group B are thus strictly worse-off, while none improve. Without the subsidy offering, group B members had been perfectly classified, but now there exist some candidates who are mistakenly excluded. Further, one can show that candidates who are positively classified must pay more to manipulate to the new threshold in spite of receiving the subsidy benefit. This increased cost is due to the fact that the higher classification threshold imposes greater burdens on manipulation than the β subsidy alleviates.

Group A candidates are also strictly worse-off since the threshold increase eliminates false positive benefits that some members had previously been granted in the no-subsidy regime. Moreover, all candidates who manipulate must expend more to do so, since these candidates do not receive a subsidy payment. Only the learner is strictly better off with the implementation of this subsidy plan.

Additional examples in the appendix show cases in which both groups experience diminished welfare when they bear linear costs. Even when the learner has an error function that penalizes false negatives twice as harshly as false positives and thus is explicitly concerned with mistakenly excluding group B candidates, an equilibrium subsidy strategy can still make both groups worse-off.

We thus highlight two consequences of subsidy interventions: On the one hand, with reduced cost burdens, more candidates from the disadvantaged group should be able to manipulate to reach a positive classification. However, subsidy payments also allow a learner to select a classifier that is at least as strict as the one issued without offering subsidies. These are opposing forces, and these examples show that without needing to distort underlying group probability distributions or the learner's penalty function in extreme ways, the effect of mitigating manipulation costs may be outweighed by the overall impact of a stricter classifier.

This result can also be extended to show that a setup in which candidates are unable to manipulate their features at all can be preferred by all three parties—groups A and B as well as the learner—to both the manipulation and subsidy regimes. We provide an informal statement of this proposition below and defer the interested reader to its formal statement and demonstration in the appendix.

Proposition 5. *There exist general cost functions such that the outcomes issued by a learner's equilibrium classifier under a non-manipulation regime is preferred by all parties—the learner, group A, and group B—to outcomes that arise both under her equilibrium manipulation classifier and under her equilibrium subsidy strategy.*

5 DISCUSSION

Social stratification is constituted by forms of privilege that exist along many different axes, weaving and overlapping to create an elaborate mesh of power relations. While our model of strategic manipulation does not attempt to capture this irreducible complexity, we believe this work highlights a likely consequence of the expansion of algorithmic decision-making in a world that is marked by deep social inequalities. We demonstrate that the design of classification systems can grant undue rewards to those who *appear* more meritorious under a particular conception of merit while justifying exclusions of those who have failed to meet those standards. These consequences serve to exacerbate existing inequalities.

Our work also shows that attempts to resolve these negative social repercussions of classification, such as implementing policies that help disadvantaged populations manipulate their features more easily, may actually have the opposite effect. A learner who has offered to mitigate the costs facing these candidates may be encouraged to set a higher classification standard, underestimating the deeper disadvantages that a group encounters, and thus serving to further exclude these populations. However, it is important to note that these unintended consequences do not always arise. A conscientious learner who offers subsidies to equalize the playing field can guard against such paradoxes by making sure to classify agents in the same way even when offering to mitigate costs.

Other research in signaling and strategic classification has considered models in which manipulation is desirable from the learner's point of view [12, 20]. Though this perspective diverges from the one we consider here, we acknowledge that there do exist cases in which manipulation serves to improve a candidate's quality and thus leads a learner to encourage such behaviors. It is important to note, however, that although this account may accurately represent some social classification scenarios, differential group access to manipulation remains an issue, and in fact, cases in which manipulation genuinely improves candidate quality may present even more problematic scenarios for machine learning systems. As work in algorithmic fairness has shown, feedback effects of classification can lead to deepening inequalities that become "justified" on the basis of features both manipulated and "natural" [9].

The rapid adoption of algorithmic tools in social spheres calls for a range of perspectives and approaches that can address a variety of domain-specific concerns. Expertise from other disciplines ought to be imported into machine learning, informing and infusing our research in motivation, application, and technical content. As such, our work seeks to investigate, from a theoretical learning perspective, some of the potential adverse effects of what sociology has called "quantification," a world increasingly governed by metrics. In doing so, we bring in techniques from game theory and information economics to model the interaction between a classifier and its subjects. This paper adopts a framework that tries to capture the genuine unfair aspects of our social reality by modeling group inequality in a population of agents. Although this perspective deviates from standard idealized settings of learner-agent interaction, we believe that so long as machine learning tools are designed for deployment in the imperfect social world, pursuing algorithmic fairness will require us to explicitly build models and theory to address critical issues such as social stratification and unequal access.

ACKNOWLEDGEMENTS

We thank Alex Frankel, Rupert Freeman, Manish Raghavan, Hanna Wallach, and Glen Weyl for constructive input and discussion on this project and related topics.

REFERENCES

[1] Emrah Akyol, Cedric Langbort, and Tamer Basar. 2016. Price of transparency in strategic machine learning. (2016). CoRR arXiv:1610.08210.
[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).
[3] Peter Auer and Nicolo Cesa-Bianchi. 1998. On-line learning with malicious noise and the closure algorithm. *Annals of mathematics and artificial intelligence* 23, 1-2 (1998), 83–99.
[4] Wolfgang Ballwieser, G Bamberg, MJ Beckmann, H Bester, M Blickle, R Ewert, G Feichtinger, V Firchau, F Fricke, H Funke, et al. 2012. *Agency theory, information, and incentives.* Springer Science & Business Media.
[5] Michael Brückner and Tobias Scheffer. 2011. Stackelberg games for adversarial prediction problems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*
[6] David Card and Jesse Rothstein. 2007. Racial segregation and the black–white test score gap. *Journal of Public Economics* 91, 11–12 (2007), 2158–2184.
[7] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Proxy non-discrimination in data-driven systems. (2017). CoRR arXiv:1707.08120.
[8] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. 2018. Strategic Classification from Revealed Preferences. In *Proceedings of the ACM Conference on Economics and Computation.*
[9] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the Conference on Fairness, Accountability and Transparency.*
[10] Joan Esteban and Debraj Ray. 2006. Inequality, lobbying, and resource allocation. *American Economic Review* 96, 1 (2006), 257–279.
[11] Virginia Eubanks. 2018. *Automating inequality: How High-tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press.
[12] Alex Frankel and Navin Kartik. Forthcoming, 2018. Muddled information. *Journal of Political Economy* (Forthcoming, 2018).
[13] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence.*
[14] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science.*
[15] Kory D Johnson, Dean P Foster, and Robert A Stine. 2016. Impartial predictive modeling: Ensuring fairness in arbitrary models. (2016). CoRR arXiv:1608.00528.
[16] Michael Kearns and Ming Li. 1993. Learning in the presence of malicious errors. *SIAM J. Comput.* 22, 4 (1993), 807–837.
[17] Andrew Kephart and Vincent Conitzer. 2015. Complexity of Mechanism Design with Signaling Costs. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems.*
[18] Andrew Kephart and Vincent Conitzer. 2016. The revelation principle for mechanism design with reporting costs. In *Proceedings of the ACM Conference on Economics and Computation.*
[19] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems.*
[20] Jon Kleinberg and Manish Raghavan. 2018. How Do Classifiers Induce Agents To Invest Effort Strategically? (2018). CoRR arXiv:1807.05307.
[21] Jean-Jacques Laffont and David Martimort. 2009. *The Theory of Incentives: The Principal-Agent Model.* Princeton University Press.
[22] Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. Forthcoming, 2019. The Social Cost of Strategic Classification. *Conference on Fairness, Accountability, and Transparency.*
[23] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Broadway Books.
[24] Bilal Qureshi, Faisal Kamiran, Asim Karim, and Salvatore Ruggieri. 2016. Causal Discrimination Discovery Through Propensity Score Analysis. (2016). CoRR arXiv:1608.03735.
[25] Michael Spence. 1978. Job market signaling. In *Uncertainty in Economics.* 281–306.
[26] Klaus Spremann. 1987. Agent and principal. In *Agency theory, information, and incentives.* Springer, 3–37.
[27] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.