Group-Fair Classification with Strategic Agents

Andrew Estornell
Washington University in Saint Louis

Sanmay Das
George Mason University

Yang Liu
University of California Santa Cruz

Yevgeniy Vorobeychik
Washington University in Saint Louis

ABSTRACT

The use of algorithmic decision making systems in domains which impact the financial, social, and political well-being of people has created a demand for these to be "fair" under some accepted notion of equity. This demand has in turn inspired a large body of work focused on the development of fair learning algorithms which are then used in lieu of their conventional counterparts. Most analysis of such fair algorithms proceeds from the assumption that the people affected by the algorithmic decisions are represented as immutable feature vectors. However, strategic agents may possess both the ability and the incentive to manipulate this observed feature vector in order to attain a more favorable outcome. We explore the impact that strategic agent behavior can have on group-fair classification. We find that in many settings strategic behavior can lead to *fairness reversal*, with a conventional classifier exhibiting higher fairness than a classifier trained to satisfy group fairness. Further, we show that fairness reversal occurs as a result of a group-fair classifier becoming more *selective*, achieving fairness largely by excluding individuals from the advantaged group. In contrast, if group fairness is achieved by the classifier becoming more *inclusive*, fairness reversal does not occur.

ACM Reference Format:
Andrew Estornell, Sanmay Das, Yang Liu, and Yevgeniy Vorobeychik. 2023. Group-Fair Classification with Strategic Agents. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3593013.3594006

1 INTRODUCTION

The increasing deployment of algorithmic decision making systems in social, political, and economic domains has brought with it a demand that fairness of decisions be a central part of algorithm design. While the specific notion of fairness appropriate to a domain is often a matter of debate, several have come to be commonly used in prior literature, such as positive (or selection) rate and false positive rate. A common goal in the design of fairness-aware (*group-fair*) algorithms is to balance predictive efficacy (such as accuracy) with achieving near-equality on a chosen fairness measure among demographic categories, such as race or gender. A question that arises in many domains where such "fair" algorithms could be used is whether they are susceptible to, and create incentives for, manipulation by agents who may *misrepresent* themselves in order to achieve better outcomes. For example, in selection of individuals to receive assistance from social service programs or selection of individuals for loans, it may be possible for applicants to misreport the number of dependents, income, or other self-reported characteristics, and, in some cases, even the sensitive attribute itself.

We investigate the effects of such strategic manipulation of a binary *group-fair* classifier. In the social services example, the classifier may decide whether an applicant receives assistance, and the fairness criterion could be approximate equality of selection rate between male and female applicants. First, we observe that the ability of individuals to manipulate the features a classifier uses can lead to *fairness reversal*, with the conventional (accuracy-maximizing) classifier exhibiting greater fairness than a group-fair classifier. We demonstrate this phenomenon on several standard benchmark datasets commonly used in evaluating group-fair classifiers. Next, we theoretically investigate conditions under which fairness reversal occurs. We prove that the key characteristic that leads to fairness reversal is that the group fair classifier becomes more selective, excluding some of the individuals in the advantaged group from being selected. Moreover, we show that this condition is sufficient for fairness reversal for several classes of functions measuring feature misreporting costs. In contrast, we experimentally demonstrate that when a group-fair classifier exhibits inclusiveness instead by selecting additional individuals from the disadvantaged group, fairness reversal does not occur.

Summary of results: We begin by observing empirically the phenomenon of fairness reversal, exhibited on a number of datasets commonly used in benchmarking group-fair classification efficacy. The key factor that results in fairness reversal is the extent to which group fairness is achieved through increased selectivity (the fair classifier f_F positively classifies fewer inputs than the conventional classifier f_C) as opposed to increased inclusiveness (f_F positively classifies more inputs than f_C). Next, we examine this issue theoretically, and prove that selectivity is a sufficient condition for fairness reversal. Further, we show that, under some additional conditions, selectivity is also a necessary condition. These results obtain for two common classes of functions measuring the cost of misreporting attributes, and explain our empirical observations.

2 RELATED WORK

Our work is closely related to two major strands in the literature: algorithmic group-fair learning and adversarial, or strategic, learning.

The *algorithmic fairness* literature aims to study the extent to which algorithmic decisions are perceived as unfair, for example, by being inequitable to historically disadvantaged groups [2, 4, 5, 10].

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594006

Many approaches have been introduced, particularly in machine learning, that investigate how to balance fairness and task-related efficacy, such as accuracy [1, 8, 16, 20, 24, 27, 43–45]. Many of these impose hard constraints to ensure that pre-defined groups are near-equitable on some exogenously specified metric, e.g., selection (positive) rate [1, 24, 44], although alternative means, such as modifying the data to eliminate disparities, have also been proposed [9, 16].

Within the domain of algorithmic fairness, our work is related to recent investigations into the effects of distribution shift, or data mismeasurement, on fair learning [17, 33, 37, 38]. These works asses the efficacy of fair learning in settings in which data is noisy, or settings in which training data and testing data are sourced from separate distributions. This line of research considers worst-case, or random, distribution shifts, which is distinct from our setting in which we explicitly consider shifts caused by strategic agent behavior. Moreover, these works compare model fairness and performance under distribution shifts or noise with model fairness and performance under no distribution shifts or noise. This is contrast to our work which examines fairness and performance or a particular (fair) model, against an alternate choice of (fairness-agnostic) model.

The *adversarial learning* literature, initially motivated by security considerations, such as malware detection [22, 30, 41], has come to have a far broader scope, including social applications [3, 6, 12, 19]. In the latter context, this is known as *strategic classification*, to indicate the concern that individuals impacted by algorithmic decisions change their features. In most cases, the strategic aspect here is actual *misreporting* of features, which is our concern. However, a related but distinct, line of work considers individuals who actually *change their features* (rather than misreport these) to achieve a better outcome [7, 25]. A broader related area of *performative prediction* considers more general changes in behavior induced by algorithmic systems [33, 34]. The intersection between strategic classification and fairness is particularly salient to our work, and has featured studies that highlight the inequity that results from strategic behavior by individuals [21], as well as social cost disparities resulting from making classifiers robust to strategic behavior [32, 43]. Our goal, however, is quite distinct: we investigate the extent to which *group-fair* learning *itself* leads to greater inequity than non-group-fair baselines due to strategic manipulation of features. Finally, Liu et al. [29] consider a closely related issue of fairness reversal that may result from a population adapting to a classifier. However, their analysis is at the *population* level, assuming known prediction scores; in contrast, we delve into individual-level manipulation of features, and build results using popular agency models.

3 PRELIMINARIES

We consider a setting with a population of agents, each characterized by a feature vector $\mathbf{x} \in \mathcal{X}$, a group $g \in G \equiv \{0, 1\}$ to which they belong (as is common in much prior literature, we treat groups as binary), and a (true) binary label $y \in \mathcal{Y} \equiv \{0, 1\}$, denoting, for example, the agent's qualification (for a service, employment, bail, etc). Let $\mathcal{D}$ be the joint distribution over $G \times \mathcal{X} \times \mathcal{Y}$. We define $p(\mathbf{x})$ as the *marginal* pdf of $\mathbf{x}$, and assume that $p(\mathbf{x}) > 0$ for each $\mathbf{x} \in \mathcal{X}$.

Since using the sensitive group membership feature may pose a legal challenge, we assume that neither the conventional nor the group-fair classifier do so at prediction time (but may at training time); from a analytical perspective group-aware classifiers (those that use group membership at prediction time) are equivalent to group-unaware classifiers from the perspective of agent manipulations, so long as group membership can be misreported in a similar fashion to other features. As such we provide a set of empirical results demonstrating that fairness reversals occur for group aware classifiers as well, but defer discussion of group-aware results to Section E.3 of the Supplement. We denote the conventional classifier by f_C, while the group-fair classifier is denoted by f_F, and both map from the domain of features $\mathcal{X}$ to the set of binary labels $\mathcal{Y}$. Let $\mathcal{M}(f; g)$ be a measure of efficacy (e.g., positive rate) of f restricted to a group g, and define

$$U(f; \mathcal{M}) = \left|\mathcal{M}(f|g = 1) - \mathcal{M}(f|g = 0)\right|.$$

We shorten this notation to $U(f)$ where $\mathcal{M}$ is clear from context. We assume that the conventional classifier aims to maximize accuracy, i.e., $f_C = \arg\max_f \mathbb{P}_{(\mathbf{x},y)}\big(f(\mathbf{x}) = y\big)$, while f_F aims to balance accuracy and fairness, solving

$$f_F = \arg\max_f\ (1 - \alpha)\mathbb{P}_{(\mathbf{x},y)}\big(f(\mathbf{x}) = y\big) - \alpha U(f; \mathcal{M}),$$

where $\alpha \in [0, 1]$ specifies the relative weight of accuracy and fairness terms.

In the literature fairness is sometimes defined with hard constraints, rather than the soft constraints of α-fairness, for example

$$f_F = \arg\max_f\ \mathbb{P}_{(\mathbf{x},y)}\big(f(\mathbf{x}) = y\big) \quad \text{s.t. } U(f; \mathcal{M}) \leq \beta.$$

In general these two formulations are not equivalent, however in the cases we study (PR, FPR, and TPR fairness) soft constrained and hard constrained fairness are equivalent, in the sense that for any α there exists a β such that the classifiers produced under either formulation are equivalent, and viceversa. This is given more precisely as Lemma A.4 in the Supplement. As such our results hold for either case; we elect to study the problem through the lens of α-fairness simply for ease of presentation.

We consider the impact of strategic behavior of agents when they face a classifier f (whether conventional or group-fair). Specifically, we suppose that each agent with features $\mathbf{x}$ can modify these, transforming them into another feature vector $\mathbf{x}'$ that is reported to the classifier. In doing so, the agent incurs a cost, captured by a manipulation cost function $c(\mathbf{x}, \mathbf{x}') \geq 0$ [18, 19, 30]. Cost functions are assumed to be bounded [1] over the domain $\mathcal{X} \times \mathcal{X}$.

We study two common families of manipulation cost functions:

Feature-monotonic costs: Manipulation cost $c(\mathbf{x}, \mathbf{x}')$ is monotonic in $||\mathbf{x} - \mathbf{x}'||$ (larger manipulations are more costly).

Outcome-monotonic costs: Manipulation cost $c(\mathbf{x}, \mathbf{x}')$ is monotonic in $\mathbb{P}(y = 1|\mathbf{x}') - \mathbb{P}(y = 1|\mathbf{x})$ where $c(\mathbf{x}, \mathbf{x}') = 0$ for any $\mathbf{x}'$ such that $\mathbb{P}(y = 1|\mathbf{x}) > \mathbb{P}(y = 1|\mathbf{x}')$ (manipulations leading to better outcomes are more costly).

[1] Boundedness of c is a rather mild assumption and holds for any continuous function when $\mathcal{X}$ is a closed and bounded set, e.g. $[0, 1]^d$. This assumption is used primarily to avoid degenerated settings such as those in which no agents can manipulate, e.g., $c(\mathbf{x}, \mathbf{x}') = \infty$ for all $\mathbf{x} \neq \mathbf{x}'$.

For example, if the problem domain involves lending, feature-monotonic costs can correspond to the mental and physiological burden of dishonesty [39], or to the likelihood of failing an authenticity check [14], while outcome-monotonic costs can correspond to the required time investment to identify a productive manipulation, or the likelihood of being audited [15] (applications more likely to succeed are also more likely to be audited) and incurring associated penalties.

We define the agent's utility as

$$u(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}') - f(\mathbf{x}) - 1/B \cdot c(\mathbf{x}, \mathbf{x}'),$$

where B is a parameter trading off costs and benefits of manipulation. Following the standard setting in strategic classification or adversarial machine learning, we assume any misreporting behavior would not change the true label y associated with $\mathbf{x}$. We assume that all agents are rational utility maximizers. Thus, since $f(\mathbf{x}') - f(\mathbf{x}) \leq 1$, the agent will misreport its features only when $c(\mathbf{x}, \mathbf{x}') \leq B$. Additionally, the agent will not misreport if $f(\mathbf{x}) = 1$ (they are selected even when truthfully reporting $\mathbf{x}$). Consequently, we can equivalently view B as an upper bound on the costs that agents are willing to incur from misreporting their features, that is, the *manipulation budget*.

We next formalize the notion of a fairness reversal in the presence of strategic agents (i.e., what it means for strategic agent behavior to lead to f_F becoming less fair than f_C).

Definition 3.1. ***(Fairness Reversal)*** *Let $\mathcal{M}$ be a measure of efficacy, f_F be a classifier which is* group-fair *with respect to $U(f; \mathcal{M})$ and f_C be a conventional accuracy-maximizing classifier. Suppose that $U(f_F; \mathcal{M}) < U(f_C; \mathcal{M})$. Let $f_C^{(c,B)}, f_F^{(c,B)}$ be the induced classifiers when agents best respond to f_C and f_F respectively with manipulation cost $c(\mathbf{x}, \mathbf{x}')$ and budget B. We say that a* budget B leads to fairness reversal *between f_C and f_F if $U(f_F^{(c,B)}; \mathcal{M}) \geq U(f_C^{(c,B)}; \mathcal{M})$.*

We will then say that fairness reversal between f_F and f_C occurs if there is some strategic manipulation budget B which leads to fairness reversal, that is, for this budget, f_C becomes more fair than f_F after manipulation. Note that if the budget B is 0, f_F will be more fair than f_C by construction, whereas if the budget is infinite, as long as any input is classified as the positive class, all individuals can misreport their features to be this class, and consequently both classifiers are fair in the sense that every input is predicted as 1. As a result, our analysis is focused solely on the intermediate cases between these extremes.

4 FAIRNESS REVERSAL

Our central goal is to understand the conditions under which *fairness reversal* occurs in strategic settings, that is, when a fair classifier f_F becomes less fair than its conventional counterpart f_C if agents act strategically. Fairness reversal occurs when there is a range of strategic manipulation budgets B for which the conventional classifier f_C exhibits greater fairness than the group-fair model f_F. In this section, we study this phenomenon empirically, demonstrating that it is commonly observed for several benchmark datasets.

Datasets and Algorithms. For our empirical study, we use five datasets commonly used as benchmarks for group-fair classification: **Adult:** Dataset of working professionals where the goal is to predict high or low income (protected feature: gender) [13, 26]. **Community Crime:** Dataset of communities where the objective is to predict if the community has high crime (protected feature: race) [13, 36]. **Law School:** Dataset of law students where the objective is to predict bar-exam passage (protected feature: race) [42]. **Student:** Dataset of students where the objective is to predict a student receiving high math grades (protected feature: race) [11, 13]. **Credit:** Dataset of people applying for credit where the objective is to predict creditworthiness (protected feature: age) [13].

All five datasets have binary outcomes, and we label the more desirable outcome for the individuals by $y = 1$ (e.g., having a high income in the Adult dataset), with the less desirable outcome labeled by $y = 0$. Consequently, higher *positive rate (PR), true positive rate (TPR),* or *false positive rate (FPR)* is more desirable for individuals. Group membership in each dataset is determined by race, gender, or age which in these datasets corresponds to a binary feature (as in [23] the age feature is made binary by considering those older than 25 as Old, and those 25 or younger as Young). A detailed breakdown of the datasets can be found in Section E.7 of the Supplement. In all cases, we refer to the "advantaged" group (e.g. the group with higher *PR* for *PR* based fairness) as group 1, or G_1, while the disadvantaged group is referred to as 0 or G_0. In our experiments, we only consider features that can potentially be manipulated (see Section E.7 of the Supplement for further details). We use four types of *conventional* classifiers for f_C, namely logistic regression (LGR), support vector machines with an RBF kernel (SVM), neural networks (NN), and gradient boosting trees (GB), and three group-fair approaches to obtain f_F, *Reductions* [1], *GerryFair* [24], and *EqOdds* [35]. The first two are inprocessing methods which learn a fair model direction on a given dataset, while the third remedies unfairness through postprocessing the predictions of a conventional classifier. To study strategic manipulation, we use a mix of local search for categorical features [28, 40] and projected gradient descent (PGD) for continuous features [31]; further details are provided in Section E.6 of the Supplement.

Fairness reversals under strategic agent behavior. In Figure 1 we investigate fairness reversals on three datasets with both Reductions and EqOdds fairness methods; additional experiments in Section E of the Supplement show that this illustration is representative in the sense that although fairness reversals do not occur in all cases, they are quite common. Consider first Figure 1 (top), which examines settings where predictions do not take the sensitive features as an input (we call these *group-agnostic* classifiers). In these three plots, the dashed line corresponds to f_C, and the rest are group-fair classifiers f_F for different values of α (recall that higher α entails greater importance of group fairness). What we observe is that in many cases, particularly when α is not very high, there is a range of budget values B for which f_F becomes less fair than f_C. Moreover, in many cases, this range is considerable. In Figure 1 (bottom plots), where group-fair classifiers are *group-aware*, including the sensitive feature as an input, the fairness reversal phenomenon is even more dramatic (not that EqOdds attempts to achieve 0 unfairness between groups, i.e., $\beta = 0$ is used in all experiments instead of α)

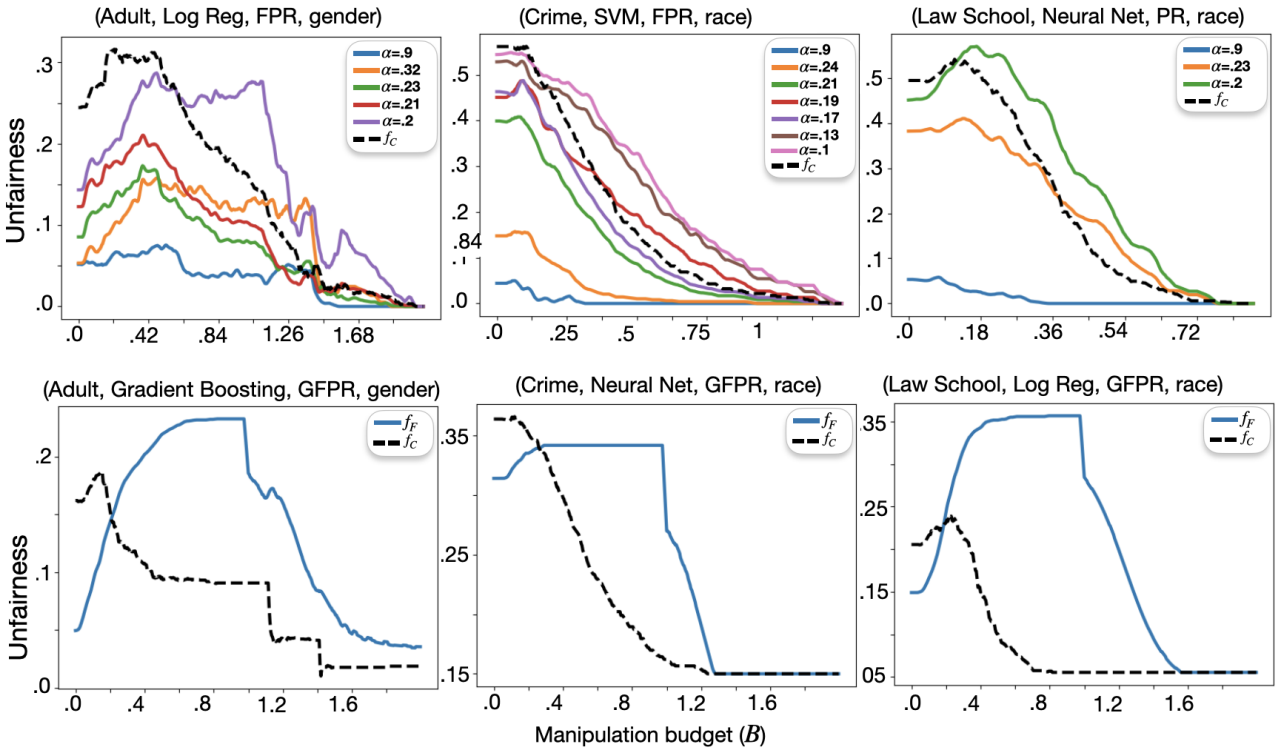

Figure 1: Difference in unfairness between groups on several datasets as a function of the manipulation budget B when manipulation cost is $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2$. The dashed black lines correspond to f_C and colored lines correspond to f_F. Fairness reversal occurs when one of the colored lines is above the black line. The top row displays results when f_F is learned via the Reductions algorithm, with fairness defined in terms of PR, TPR, or FPR, for several different values of α. The bottom row displays results when f_F is learned via the EqOdds algorithm, with fairness defined in terms of generalized false positive rate GFPR (i.e. expected FPR: Definition 1 of [35]). Reductions is group-agnostic, and EqOdds is group-aware.

. In this experiment, when best responding agents are capable of misreporting their group as if it where a feature in $\mathbf{x}$ (fairness is still computed with true group membership). Due to the particular nature of EqOdds, specifically its handling of agents from different groups, we observe a sharp change in fairness at $B = 1$, the precise budget for which misreporting group membership is feasible.

Figure 1 exhibits several additional phenomena. Note, in particular, that in many cases the unfairness (i.e., FPR difference between the groups) initially *increases* as the budget increases, but in all cases as budgets B keep increasing, eventually unfairness vanishes *as a result of strategic behavior by agents*. Furthermore, much as we observe this initial unfairness increase for both f_C and f_F, it appears *amplified* for some of the group fair classifiers f_F.

What causes fairness reversal? As we formally prove below, the essential condition is *selectivity* of fair classifier f_F compared to f_C. Specifically, in binary classification, there are, roughly, two ways one can improve fairness on a given dataset (that is, without any consideration of strategic behavior); either through *inclusiveness* (positively classifying additional agents from the disadvantaged group by changing their predicted class to 1), or through *selectivity* (negative classifying some of the members of the advantaged group by changing their predicted class 1 to 0).

Our key observation is that *selectivity* leads to fairness reversals, while *inclusiveness* does not. Specifically, we observe that as the number of agents positively classified under f_C, but negatively under f_F, is larger than the number of agents negatively classified by f_C, but positively under f_F, fairness reversals are more commons.

We illustrate this in Figure 2, which shows the decision boundaries of f_F and f_C (top row), as well as associated fairness as a function of budget (bottom row) for several combinations of dataset, classifier, and fairness definition. On the Adult and Crime datasets (first two columns), fairness is achieved predominantly through selectivity, as the orange region (f_C) includes few additional green points (disadvantaged group) compared to the blue region (f_C), but excludes many blue points (advantaged group). This is given more precisely in terms of the respective group-wise positive rates for f_C and f_F; in the first two examples the positive rates on both groups drops when switching from f_C to f_F, while in the third case the positive rate for both groups increases. This, in turn, leads to instances of fairness reversal (bottom row first column). In the Law School dataset (third column), in contrast, fairness is achieved primarily through inclusiveness, and f_F remains more fair than f_C over a broad range of strategic manipulation budgets B. The

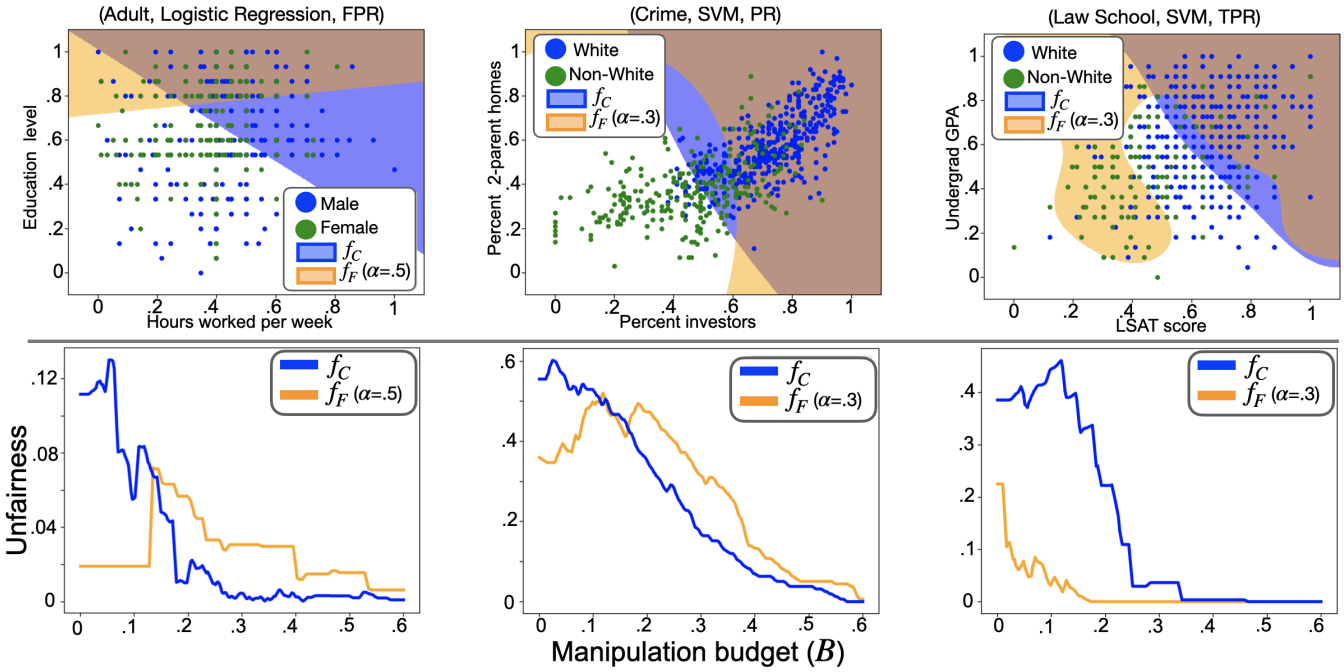

Figure 2: Fairness reversals and selectivity of classifiers on two ordinal features. The top row shows regions with positive predictions (blue for f_C and light orange for f_F) using two features (corresponding to the axes), and dot colors correspond to the sensitive demographics. The darker orange region corresponds to an overlap between the positive predictions of f_C and f_F. The bottom row shows the relative unfairness between demographic groups (for the classifiers shown in the top row) as a function of strategic manipulation budget B (lower means more fair). In the top row, the fraction of each group being positively classified under f_C is, Adult: (Male: .63, Female: .45), Crime: (White: .84, None-White: .26), Law: (White: .64: None-White: .35) alternatively under f_F is, Adult: (Male: .42, Female: .39), Crime: (White: .62, None-White: .23), Law: (White: .62: None-White: .51)

reason that selectivity leads to fairness reversal is that those from the advantaged group who are excluded tend as a result to be closer to the decision boundary than those from the disadvantaged group. In Section E.1 of the Supplement we provide further results linking selectivity of the fair classifier to fairness reversals. In this section we also observe that when strategic agent behavior (for some manipulation budget) results in a fairness reversal between f_F and f_C, the relative accuracy of the classifiers is also reversed (for some potentially different manipulation budget), implying a fundamental relationship between fairness and accuracy when agents are strategic.

Unfairness of f_F. Lastly we remark on the relationship of the between the manipulation budge B and the unfairness of the fair classifier f_F. As seen in Figures 2 and 1, the unfairness of f_F is frequently increasing in B (for small values of B). To provide insight into this phenomenon we look to the case of single variable prediction as showing in Figure 3. This figure shows the error and unfairness of a single variable classifier (i.e., a threshold classifier with threshold θ) when using a student's LSAT score to predict whether they will pass the bar exam. Since manipulations change model decisions only in a single direction (negative predictions become positive), predicting on strategically altered data amounts to predicting on unaltered data with a lower threshold . As the manipulation budget B grows, the corresponding threshold becomes increasingly smaller. Thus, when f_F is more selective than f_C, i.e. $\theta_F > \theta_C = 0.57$, the unfairness of f_F will initially increase as B increases. In the case of multivariate prediction, the increased unfairness of f_F stems from a similar

Next, we study fairness and accuracy reversals in strategic classification settings theoretically, demonstrating that selectivity is indeed a sufficient (and, under some additional qualifications, necessary) condition for fairness reversal.

5 THEORETICAL ANALYSIS

In this section we provide theoretical explanations of the empirical observations made in the previous section. We start with single-variable classifiers and then proceed to generalize our observations to multi-feature classifiers. Our key finding is that *selectivity* (defined next) is in fact a sufficient condition for fairness reversal, providing a theoretical underpinning for the empirical observations above. Additionally, we investigate the underlying *causes* of fair classifiers becoming more selective, and provide conditions on the underlying distribution for this to be the case. In the cases of single variable classifiers with feature-monotonic costs and multivariable classifiers with outcome-monotonic costs, we further demonstrate that selectivity also leads to *accuracy reversals* (strategic behavior causes the fair classifier to become *more* accurate than the conventional model), and outline conditions on the underlying distribution

such that selectivity is not just sufficient, but also necessary for both of these phenomena. When strategic agent behavior results in both a fairness and accuracy reversal, the functionality of both classifiers has fundamentally swapped; the accuracy driven (conventional) model f_C is no longer the most accurate model and the fairness driven (fair) model f_F is no longer the most fair model. Prior to our results, we first formalize the notion of classifier selectivity.

Definition 5.1. *Let* $\mathcal{X}_{f_C} = \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}$ *and* $\mathcal{X}_{f_F} = \{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\}$. *We say that* f_F is more selective than f_C *if* $\mathcal{X}_{f_F} \subset \mathcal{X}_{f_C}$.

That is, f_F is more selective than f_C if the set of positively classified examples under f_F is a subset of those positively classified under f_C. While this definition of selectivity is slightly more restrictive than the type of selectivity found in our empirical results, the subset propriety is a driving force behind the fairness reversals observed in practice. Selectivity can be interpreted as the fair model f_F, achieving fairness by "excluding" additional agents from positive classification, compared to f_C. As an example, under PR-based fairness let G_0 be the group with lower PR and G_1 be the group with higher PR under f_C (TPR and FPR hold similarly). A model designer could improve the fairness of f_C by positively classifying more agents in G_0 or negatively classifying more agents in G_1 (or a combination of both). In the latter case, members of G_0 are "excluded" from positive classification, and the resulting model is considered to be *more selective*. Note that this type of exclusion is precisely the means through which fairness is achieved in Figure 2 (center).

5.1 Single Variable Classifier

We begin our theoretical exploration of fairness reversals with an exemplar case: a single variable threshold classifier. In this setting agents possess a single ordinal feature x. For simplicity we demonstrate our results for a continuous feature $x \in [0, 1]$, but the results hold for any ordinal feature (discrete or continuous) . Both the conventional classifier (selected for maximal accuracy) and fair classifier (selected for a weighted combination of accuracy and fairness with respect to a fairness metric M) can be expressed as a single parameter $\theta_C, \theta_F \in [0, 1]$ respectively where $f(x) = \mathbb{I}[x \geq \theta]$.

Our first result is that in single-feature classification, higher selectivity of the group-fair classifier (i.e. $\theta_C < \theta_F$) is sufficient for fairness reversal.

Theorem 5.2. *Suppose fairness is defined by PR, TPR, or FPR,* $c(x, x')$ *is monotone in* $|x' - x|$, θ_C *is the most accurate, and* θ_F *the optimal* α*-fair, threshold. If* $\theta_C < \theta_F$, *then there exists a budget* B *that leads to fairness reversal between* f_F *and* f_C.

Proof Sketch. The full proof is provided in Section B of the Supplement. Here we provide a proof sketch for continuous c, a similar line of reasoning, with a few additional edge cases, holds for discontinuous c. The unfairness of threshold θ w.r.t. to the distribution $\mathcal{D}$ and fairness metric $\mathcal{M} \in \{\text{PR}, \text{TPR}, \text{FPR}\}$ is expressed as,

$$U_{\mathcal{D}}(\theta) = \left|\mathcal{M}_{\mathcal{D}}(\theta|g = 1) - \mathcal{M}_{\mathcal{D}}(\theta|g = 0)\right|,$$

For a given threshold θ and manipulation budget B the best response of an agent with true feature x is

$$x_\theta^{(B)} = \arg\max_{x'} \left(\mathbb{I}[x' \geq \theta] - \mathbb{I}[x \geq \theta]\right) \quad \text{s.t. } c(x, x') \leq B,$$

When agents from $\mathcal{D}$ play this optimal response, let the resulting distribution be $\mathcal{D}_\theta^{(c,B)}$. The difference in unfairness between classifiers when agents are strategic is $U_{\mathcal{D}_{\theta_C}^{(c,B)}}(\theta_C) - U_{\mathcal{D}_{\theta_F}^{(c,B)}}(\theta_F)$. Since both f_C and f_F are thresholds, and c is feature-monotonic, the decisions of θ_C, θ_F on the modified distribution $\mathcal{D}_\theta^{(c,B)}$ can be expressed as decisions of modified thresholds $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ on the original distribution $\mathcal{D}$, i.e.,

$$U_{\mathcal{D}_{\theta_C}^{(c,B)}}(\theta_C) - U_{\mathcal{D}_{\theta_F}^{(c,B)}}(\theta_F) = U_{\mathcal{D}}(\theta_C^{(c,B)}) - U_{\mathcal{D}}(\theta_F^{(c,B)})$$

where

$$\theta_C^{(c,B)} = \text{argmin}_x x \text{ s.t. } c(x, \theta_C) \leq B$$

and

$$\theta_F^{(c,B)} = \text{argmin}_x x \text{ s.t. } c(x, \theta_F) \leq B$$

Given these modified threshold, we see that strategic agent behavior results in a *lowering* of each threshold as more agents are now able to achieve positive classification; this is due to the fact that only negatively classified agents will behavior strategically, their goal being to achieve positive classification. Moreover, when considering $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ as functions of B, both are monotonically decreasing in B (due to the the monotonicity of c), and $\theta_C^{(c,B)} \leq \theta_F^{(c,B)}$ for all B (due to $\theta_C < \theta_F$).

Since fairness is defined in terms of PR, FPR, or TPR the constant function $f(x) = 1$ has unfairness 0 for any distribution. Thus, $\theta_C^{(c,B)} = 0$ implies $U_{\mathcal{D}}(\theta_C^{(c,B)}) = 0$. Let

$$B' = \sup\{B \in \mathbb{R}_+ : U_{\mathcal{D}}(\theta_C^{(c,B)}) > 0\},$$

Note that B' is guaranteed to exist due to $U_{\mathcal{D}}(\theta_C^{(c,0)}) > U_{\mathcal{D}}(\theta_F^{(c,0)}) \geq 0$ and the boundedness of $c(x, x')$. Since $U_{\mathcal{D}} \geq 0$ and c is continuous, there must exist some $\varepsilon > 0$ such that over the interval $B \in [B' - \varepsilon, B']$ the unfairness $U_{\mathcal{D}}(\theta_C^{(c,B)})$ is strictly decreasing in B. If

$$U_{\mathcal{D}}(\theta_F^{(c,B'-\varepsilon)}) \geq U_{\mathcal{D}}(\theta_C^{(c,B'-\varepsilon)}) > 0,$$

then a fairness reversal has already occurred for budget $B' - \varepsilon$, so assume otherwise. Combining the difference in relative fairness for budget $B' - \varepsilon$ with the fact that $\theta_C^{(c,B)} \leq \theta_F^{(c,B)}$ for all B, we get $\theta_C^{(c,B'-\varepsilon)} < \theta_F^{(c,B'-\varepsilon)}$. Since c is monotonic and continuous there must exist some budget $B_F > B' - \varepsilon$ such that $\theta_C^{(c,B'-\varepsilon)} = \theta_F^{(c,B_F)}$. Since $B_F \geq B' - \varepsilon$, and $U_{\mathcal{D}}(\theta_C^{(c,B)})$ is decreasing for $B \geq B' - \varepsilon$, it must be the case that

$$U_{\mathcal{D}}(\theta_C^{(c,B_F)}) = U_{\mathcal{D}}(\theta_F^{(c,B'-\varepsilon)}) \leq U_{\mathcal{D}}(\theta_F^{(c,B_F)}),$$

and a fairness reversal occurs for budget B_F. □

We now turn our attention to a complementary observation: fairness reversal is accompanied by *accuracy reversal*, that is, strategic

behavior leads to f_F having higher accuracy than f_C. This is primarily due to the fact that f_F becomes more selective and therefore more resilient to manipulation. Note that the fairness reversal and accuracy reversal need not occur for the same budget B.

Theorem 5.3. *Suppose fairness is defined by PR, TPR, or FPR, $c(x, x')$ is monotone in $|x' - x|$, θ_C is the most accurate threshold, and θ_F the optimal α-fair threshold. If $\theta_C < \theta_F$, then there exists a budget B s.t. f_F is more accurate than f_C.*

Proof Sketch. We defer the full proof to Section B of the Supplement and again give a proof sketch for continuous c. The error of threshold θ on distribution $\mathcal{D}$ is given by

$$\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{P}(\mathbb{I}[x \geq \theta] = y)$$

By the definition of θ_C, we have

$$\mathcal{L}_{\mathcal{D}}(\theta_C) \leq \mathcal{L}_{\mathcal{D}}(\theta) \quad \text{for all } \theta \in [0, 1],$$

and therefore $\mathcal{L}_{\mathcal{D}}(\theta_C) \leq \mathcal{L}_{\mathcal{D}}(\theta_F)$. Similar to the proof of Theorem 5.2, agents strategically responding to threshold classifiers θ_C, θ_F can be viewed as modified thresholds $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ operating on the true distribution $\mathcal{D}$. Both $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ are monotonically decreasing in B. Moreover, $\theta_C^{(c,B)} = 0$ implies $\mathcal{L}_{\mathcal{D}}(\theta) = \mathbb{P}(y = 0)$, since the threshold classifies all agents positively.

Let

$$B' = \sup\{B : \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B)}) < \mathbb{P}(y = 0)\},$$

i.e. B' is the "largest" manipulation budget such that the conventional threshold is not a trivial classifier (i.e., not making constant predictions) in the presence of strategic agent behavior. In a similar line of reasoning to the case of fairness reversals, there must exist some interval $[B' - \varepsilon, B']$ over which $\mathcal{L}_{\mathcal{D}}(\theta^{(c,B)})$ is strictly increasing. Again, by the fact that $\theta_C < \theta_F$, there must exist some $B_F > B' - \varepsilon$ such that $\theta_C^{(B'-\varepsilon)} = \theta_F^{(c,B_F)}$. Thus,

$$\mathcal{L}_{\mathcal{D}}(\theta_F^{(c,B_F)}) = \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B'-\varepsilon)}) \geq \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B_F)}),$$

implying that an accuracy reversal occurs for budget B_F. □

We have showed thus far that selectivity is *sufficient* for fairness and accuracy reversals, but under what conditions is it also *necessary*? Loosely speaking, when a feature x is a good predictor of both y and g, the error and unfairness of f_C and f_F are *unimodal* (defined next) with respect to the manipulation budget B.

Definition 5.4. ***(Unimodal):*** *A function $g : [a, b] \to \mathbb{R}$ is* negatively unimodal *(*positively unimodal*) on the interval $[a, b]$ if there exists an* inflection *point $r \in [a, b]$ such that f is monotone decreasing (increasing) on $[a, r]$ and monotone increasing (decreasing) on $[r, b]$.*
(All convex functions are negatively unimodal *and all concave functions are* positively unimodal*).*

Unimodality is relevant to fairness and accuracy reversals as we will see that when error is negatively unimodal and unfairness is positively unimodal, both fairness and accuracy reversals occur. We empirically demonstrate that unimodality of both functions holds frequently on real data. The condition of unimodal error and unfairness can be interpreted as both functions possessing a "sweet spot" which yields best case accuracy (or worst case unfairness). In the former, x is good predictor of true label y and in the latter x is a good predictor of g.

As an example, in Figure 3 we see this phenomenon occur on the Law School dataset when using a student's LSAT score as the predictive feature $\mathbf{x}$. Both error and unfairness (in terms of positive rate difference between groups) are both unimodal in the threshold θ. In this, we observe that LSAT score is a good predictor of both the target variable (bar passage) and the sensitive feature (race); this is a well established source of bias within this particular dataset.

We further document this relationship in Section E.4 of the Supplement and find that most ordinal features produce threshold classifiers which have (approximately) unimodal error and unfairness. In this section we also theoretically outline the precise conditions under which error and unfairness would be unimodal; these conditions essentially boil down to a correlation between $y|x$ and $g|x$, (which we observe to be the case for most ordinal features across the datasets we study). When this occurs, the selectivity of f_F is not only sufficient for fairness and accuracy reversals, but also necessary. We next formalize this in the following theorem; further details on the necessary and sufficient conditions required for fairness and accuracy reversals are provided in Section B of the Supplement.

Theorem 5.5. *Let θ_C and θ_F be the most accurate and optimal fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(\mathbf{x}, \mathbf{x}')$ is outcome monotonic, and that error (and unfairness) are negatively (positively) unimodal in θ. Then there exists a budget B such that strategic agent behavior leads to a fairness reversal if an only if f_F is more selective than F_C.*

proof sketch. When error $\mathcal{L}_{\mathcal{D}}(\theta)$ and unfairness $U_{\mathcal{D}}(\theta)$ are both unimodal in θ, the optimal conventional threshold θ_C and optimal α-fair threshold θ_F can be expressed in terms of the inflection points $x_{\mathcal{L}}$ and x_U of error and unfairness respectively. The most accurate threshold is then $\theta_C = x_{\mathcal{L}}$, and the most unfair threshold

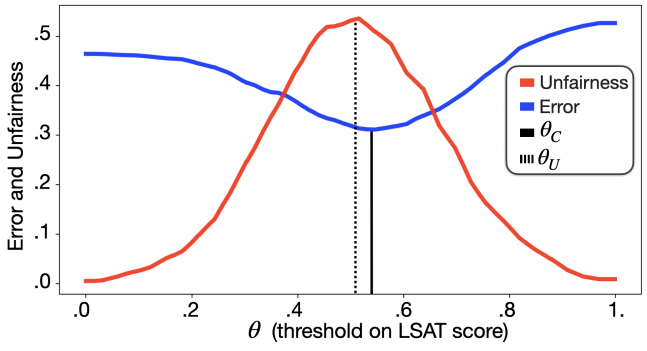

Figure 3: Error (blue) and PR-based unfairness between White and Non-White individuals (red) of a single variable classifier on the Law School dataset when using the student's LSAT score as a single predictive feature. All individuals with an LSAT score above the threshold θ are predicted positively. The thresholds θ_C and θ_U are the most accurate and least fair thresholds respectively.

is then $\theta_u = x_U$. The forward direction, i.e. when $\theta_C < \theta_F$, follows a similar of reasoning to the proof of Theorem 5.2, let $\theta_C^{(c,B)}$ and $\theta_F^{(c,B)}$ be the modified thresholds induced by agents best responding to either threshold with cost function c and budget B. Then, since $\theta_C^{(c,B)}, \theta_F^{(c,B)}$ are monotonically decreasing in B and $\theta_C < \theta_F$, there must exist a B' such that $\theta_C^{(c,B')} \leq \theta_F^{(c,B')} = \theta_C$. Thus

$$U_{\mathcal{D}}(\theta_C^{(c,B')}) \leq U_{\mathcal{D}}(\theta_C) = U_{\mathcal{D}}(\theta_F^{(c,B')})$$

and

$$\mathcal{L}_{\mathcal{D}}(\theta_F^{(c,B')}) = \mathcal{L}_{\mathcal{D}}(\theta_C) \leq \mathcal{L}_{\mathcal{D}}(\theta_C^{(c,B')}),$$

implying that a fairness and accuracy reversal occurs for budget B'.

The reverse direction, follows from the relationship between θ_F and the two inflection points θ_C, θ_U. Given the assumption that $\theta_F < \theta_C$, there are only three possible cases for the relationship between these points

$$(1)\ \theta_F < \theta_C \leq \theta_U,$$
$$(2)\ \theta_F < \theta_U \leq \theta_C,$$
$$(3)\ \theta_U < \theta_F < \theta_C$$

the strict inequalities being due to the fact that $\theta_F \neq \theta_C$ and $\theta_F \neq \theta_U$ by definition. In cases (1) and (2), no fairness or accuracy reversal can occur. Only in case (3) can a fairness or accuracy reversal occur, however we will show by contradiction that such a case is impossible.

Beginning with case (1), both error and unfairness are unimodal in $\theta_F^{(c,B)}, \theta_C^{(cB)}$, each of which is monotonically increasing in B. Since unfairness in unimodal, any $\theta \leq \theta_U$ and any $B \geq 0$ unfairness $U(\theta^{(c,B)})$ is monotonically decreasing in B. Similarly, since error is unimodal, for any $\theta \leq \theta_C$, error $\mathcal{L}(\theta^{(c,B)}$ is monotonically decreasing. Thus if $\theta_F < \theta_C$, then no accuracy reversal can occur. Similarly if $\theta_F < \theta_C \leq \theta_U$, no fairness reversal can occur, i.e. in case (1), neither reversal can occur.

In case (2) since $U_{\mathcal{D}}(\theta_F) < U_{\mathcal{D}}(\theta_C)$, and $U_{\mathcal{D}}(\theta_C^{(c,B)})$ is monotonically increasing until $\theta_C^{(c,B)} = \theta_U$, no fairness reversal will occur. Similar to case (1), $\theta_F < \theta_C$, implies that no accuracy reversal occurs either.

Thus it remains only to show that case (3) can never occur. To see this, note that for any $0 < \varepsilon < \theta_C - \theta_F$, it must be the case that both

$$U_{\mathcal{D}}(\theta_F + \varepsilon) \leq U_{\mathcal{D}}(\theta_F)$$

and

$$\mathcal{L}_{\mathcal{D}}(\theta_F + \varepsilon) \leq \mathcal{L}_{\mathcal{D}}(\theta_F)$$

Which implies that θ_F is in-fact *not* the optimal fair threshold. □

Now that we have established the critical role of selectivity in fairness reversal, we next analyze *why* that is. As mentioned previously, there are roughly two ways to achieve fairness: *inclusiveness* (classifying more examples as positive) or *selectivity* (classifying fewer examples as positive). Which of these will be the predominant outcome of training f_F depends intimately on the data distribution. We outline these conditions, as well as conditions for error and unfairness to be unimodal, via Lemmas B.4, B.5, and Theorem B in the Supplement. In particular, Theorem B provides conditions on the underlying distribution such that the optimal fair classifier will achieve fairness via selectivity. The condition in this theorem can be intuitively interpreted as follows. Suppose that S is the set of individuals selected (i.e., classified as 1) by f_C, who are also near the decision boundary of f_C. If the advantaged group (i.e., group with better average outcomes) is overrepresented in S, there is a range of parameters α such that the optimal α-fair classifier is more selective than f_C (recall that higher α places greater importance on group-fairness in learning).

5.2 General Classifiers

Next we discuss general multi-variate classifiers, generalizing several of the results from Section 5.1. First we show that when f_F is more selective than f_C, fairness reversal occurs for both feature-monotonic and outcome-monotonic cost functions. Second, we give conditions which lead to f_F being more selective than f_C. For outcome-monotonic costs, we provide two additional results: 1) greater selectivity of f_F also leads to accuracy reversal, and 2) unimodality of each classifier's error and unfairness causes selectivity to be both necessary and sufficient for fairness and accuracy reversal.

Outcome-Monotonic Costs. We begin with the case of outcome-monotonic costs. As shown by Milli et al. [32], outcome-monotonic manipulation costs result in the following best response for classifier f. Let

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x}) \quad \text{s.t. } f(\mathbf{x}) = 1.$$

If $c(\mathbf{x}, \mathbf{x}^*) \leq B$ then the best response is $\mathbf{x}' = \mathbf{x}^*$ otherwise $\mathbf{x}' = \mathbf{x}$. With this best response in hand we show that f_F having greater selectivity than f_C leads to fairness reversal.

Theorem 5.6. *Let f_C and f_F be the most accurate and optimal fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(\mathbf{x}, \mathbf{x}')$ is outcome monotonic. Then if f_F is more selective than f_C, there exists a budget B such that strategic agent behavior leads to a fairness reversal.*

Here we provide a proof sketch of this and other results; complete proofs are deferred to the Supplement (Section C).

Proof sketch. For a given classifier f, let

$$p_{\min} = \min_{\mathbf{x}: f(\mathbf{x})=1} \mathbb{P}(y = 1|\mathbf{x})$$

and let $\mathbf{x}_{\min}$ be the feature associated with $p_{\min}$ ($\mathbf{x}_{\min,C}, p_{\min,C}$ and $\mathbf{x}_{\min,F}, p_{\min,F}$ correspond to f_C and f_F respectively). When agents best respond to f the resulting manipulated classifier can be expressed as a threshold on the underlying probabilities $\mathbb{P}(y = 1|\mathbf{x})$. More specifically, let

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mathbb{P}(y = 1|\mathbf{x}) \quad \text{s.t. } c(\mathbf{x}, \mathbf{x}_{\min}) \leq B.$$

Then when agents best resound to f (inducing classifier $f^{(c,B)}$) any agent $\mathbf{x}$ with $\mathbb{P}(y = 1|\mathbf{x}) \geq \mathbb{P}(y = 1|\mathbf{x}^*)$ will be positively classified

under $f^{(c,B)}$, i.e.

$$f^{(c,B)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}(y=1|\mathbf{x}) \geq \mathbb{P}(y=1|\mathbf{x}^*) \\ 0 & \text{otherwise} \end{cases}$$

Thus $f^{(c,B)}$ can expressed as the threshold $\mathbb{P}(y=1|\mathbf{x}^*)$ operating on the conditional distribution $\mathbb{P}(y=1|\mathbf{x})$.

Since f_F is more selective than f_C, (i.e., for any $\mathbf{x} \in \mathcal{X}$, if $f_F(\mathbf{x}) = 1$ then $f_C(\mathbf{x}) = 1$), and $\mathbf{x}_{\min,F}$ is positively classified under f_F, we have,

$$f_F(\mathbf{x}_{\min,F}) = 1 = f_C(\mathbf{x}_{\min,F}) \quad \text{and therefore } \; p_{\min,C} \leq p_{\min,F}$$

Therefore the induced conventional and fair thresholds $\mathbb{P}(y=1|\mathbf{x}^*_C)$ and $\mathbb{P}(y=1|\mathbf{x}^*_F)$ acting on $\mathbf{P}(y=1|\mathbf{x})$ have the relationship that $\mathbb{P}(y=1|\mathbf{x}^*_C) \leq \mathbb{P}(y=1|\mathbf{x}^*_C)$. Thus, we see that selectivity of the fair classifier in the case of outcome-monotonic costs yields a fair threshold (on a modified distribution) which is larger than the induced conventional threshold (operating on the same distribution as the fair threshold).

While this setting is not entirely equivalent to the single variable case, the remainder of the proof follows in similar fashion to that of Theorem 5.2. In particular, the monotonicity of $\mathbb{P}(y=1|\mathbf{x}^*)$, as a function of B, implies

$$\mathbb{P}(y=1|\mathbf{x}^*_C) \leq \mathbb{P}(y=1|\mathbf{x}^*_F) \quad \text{for any } \; B,$$

which in turn implies the existence of a budget interval over which the unfairness of $f_C^{(c,B)}$ decreases below $f_F^{(c,B)}$, thus resulting in a fairness reversal. □

Similar to the single-variable case, selectivity also result in accuracy reversal.

Theorem 5.7. *Let f_C and f_F be the most accurate and optimal fair classifiers respectively. Suppose fairness is defined by PR, FPR, or TPR, and $c(\mathbf{x}, \mathbf{x}')$ is outcome-monotonic. Then if f_F is more selective than f_C, then there exists a budget B under which f_F becomes more accurate than f_C.*

Proof. The full proof (which follows from a similarly Theorem 5.3, 5.6) is deferred to Section C of the Supplement. □

Before outlining settings in which selectivity is not only sufficient but also necessary for fairness and accuracy reversals to occur, we first remark on the connection between selectivity, accuracy, and fairness. As previously noted, errors caused by strategic agent behavior are single-directional in the sense that manipulation can only induce false positive errors. As such, classifiers which are more selective are thus more robust to manipulation than their less selective counterparts. Generally speaking, this implies that for some range of manipulation budgets, a model that is more selective than the accuracy-maximizing model f_C will increase in its performative ability compared to f_C. As the performative ability of most classifiers on biased datasets is naturally tied with unfairness, the unfairness of the more robust model (more selective model) will likewise increase. Thus, we see a fundamental, albeit not necessarily universal, connection between selectivity (which in turn increases robustness) and model unfairness (which is increasing in model performance).

We next discuss unimodality in the context of outcome-monotonic costs. Empirically we observe that when costs are outcome-monotonic, the majority of classifiers tend to have error and unfairness which is (approximately) unimodal with respect to the manipulation budget B. When this occurs, selectivity of f_F becomes both necessary and sufficient.

Theorem 5.8. *Let f_C and f_F the optimal conventional and fair classifiers respectively. Suppose fairness is defined in terms of PR, TPR, or FPR fairness, and $c(x, x')$ is outcome-monotonic. When error (and unfairness) are negatively (positively) unimodal with respect to the manipulation budget B, a fairness and accuracy reversal will occur between f_F and f_C if and only if f_F is more selective than f_C (each reversal may occur at different budgets B) .*

Proof Sketch. We deffer the full proof to Section C of the Supplement. The intuition for this proof follows similarly to that of Theorem 5.5. As shown in the proof of Theorem 5.6 when agents best respond to classifier f, the decisions of f can be expressed as threshold classifier acting on the conditional probability $\mathbb{P}(y=1|\mathbf{x})$ of the original distribution $\mathcal{D}$, namely

$$f^{(c,B)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbb{P}(y=1|\mathbf{x}) \geq \mathbb{P}(y=1|\mathbf{x}^*) \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{x}^*$ is determined by the cost function c and budget B. Since $\mathbb{P}(y=1|\mathbf{x}*)$ is monotonically decreasing in B, we recover a setting similar to 5.5, in which the forward direction of the claim holds from the fact that,

$$\mathbb{P}(y=1|\mathbf{x}^*_C) \leq \mathbb{P}(y=1|\mathbf{x}^*_F), \quad \text{for all } \; B.$$

While the reverse direction holds due to the fact that when $\mathbf{x}^*_F \leq \mathbf{x}^*_C$, unfairness is monotonically decreasing for both classifiers. □

Remark 5.9. *To better contextualize unimodality of error and unfairness with respect to the manipulation budget B, we can view this condition in terms of the calibration of the score function h of the classifier f. As is typical, classifiers are defined via thresholds on their underlying score functions, i.e. $f(\mathbf{x}) = \mathbb{I}[h(\mathbf{x}) \geq \theta]$. Suppose that h is reasonably well calibrated, then for every $p \in [0,1]$, $\mathbb{P}(y=1|h(\mathbf{x}) = p) \approx p$, i.e. $h(\mathbf{x})$ is a good approximation of the conditional distribution given by $\mathbb{P}(y=1|\mathbf{x})$. When h is reasonably well calibrated, the condition that error and unfairness are unimodal w.r.t. to the manipulation budget B is equivalent to the error and unfairness of f being unimodal w.r.t. to the choice of threshold θ. Through this lens, one can see that the assumption of unimodality is likely to hold (at least approximately so) in practice as it is typically the case there is one "good" choice of threshold θ and any deviation (increasing or decreasing θ) results in strictly worse performance of f.*

Feature-Monotonic Costs. Finally, we demonstrate that selectivity remains sufficient for fairness reversal in general when costs are feature-monotonic.

Theorem 5.10. *Let f_C and f_F be the most accurate and the optimal α-fair classifier, respectively. Suppose fairness is defined by PR, FPR, or TPR and $c(\mathbf{x}, \mathbf{x}')$ is feature-monotonic. If f_F is more selective than f_C, then there exists a budget B that leads to fairness reversal between f_F and f_C.*

Proof Sketch. The full proof is deferred to Section D of the Supplement. The intuition behind this results is that trivial classifiers (i.e., those that predict $f(\mathbf{x}) = 1$ for all $\mathbf{x}$) have 0 unfairness

for PR, FPR, and TPR based fairness. As B increases, both $f_C^{(c,B)}$ and $f_F^{(c,B)}$ (the classifiers resulting from agents best responding to either classifier with budget B and cost function c) will approach 0 unfairness, not necessarily monotonically, as they become more like trivial classifiers. At some point prior to reaching trivial classification, the conventional classifier f_C will be at least as fair as f_F. This can be seen through a combination of the fact that f_F is more selective than f_C and the way in which manipulations alter the positively predicted region of a classifier when costs are feature-monotonic. In particular, f_F being more selective than f_C implies that,

$$\{\mathbf{x} \in \mathcal{X} : f_F(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C(\mathbf{x}) = 1\}.$$

Feature-monotonic cost functions preserve this subset propriety under manipulation, i.e., for any B,

$$\{\mathbf{x} \in \mathcal{X} : f_F^{(c,B)}(\mathbf{x}) = 1\} \subset \{\mathbf{x} \in \mathcal{X} : f_C^{(c,B)}(\mathbf{x}) = 1\}.$$

Thus f_F is *always* more selective than f_C, regardless of the manipulation budget B. As such, the positive rate of f_F will never exceed the positive rate of f_C, implying that $f_F^{(c,B)}$ approaches a trivial classifier more "slowly" than $f_C^{(c,B)}$, with respect to B. Moreover, prior to approaching triviality $f_F^{(c,B)}$ will effectively approach f_C, thus partially absorbing some of the original unfairness of f_C, resulting in a fairness reversal. □

Next, we provide a condition which leads f_F to be more selective than f_C. Here, we provide this condition for the PR fairness metric; analogous results for TPR and FPR are given in Section D of the Supplement. For this result, we define the following notation

$$P_{G_z} = \mathbb{P}(g = z), \quad g(\mathbf{x}) = P(g = 1|\mathbf{x})$$

and

$$\mathcal{X}_0 = \{\mathbf{x} \in \mathcal{X} : g(x) < P_{G_1} \text{ and } \mathbb{P}(y = 1|\mathbf{x}) < 1/2\}.$$

The set $\mathcal{X}_0$ represents the set of features which are less likely than chance to correspond to $g = 0$ and $y = 0$.

Theorem 5.11. *Let f_C and f_F be the most accurate and optimal α-fair classifiers respectively, and fairness defined by PR. Then f_F is more selective than f_C if and only if $0 < \alpha \leq \alpha^*$, where*

$$\alpha^* = \min_{\mathbf{x} \in \mathcal{X}_0} \frac{P_{G_0} P_{G_1} (2\mathbb{P}(y{=}1|\mathbf{x}) - 1)}{g(\mathbf{x}) + P_{G_1}\left(P_{G_1} - 2g(\mathbf{x}) - 2P_{G_1}\mathbb{P}(y{=}1|\mathbf{x})\right)}.$$

Proof. Both the conventional and fair objectives can be written as follows:

$$\begin{aligned} f_C =& \text{argmin}_f \mathbb{P}(f(\mathbf{x}) \neq y) \\ f_F =& \text{argmin}_f (1-\alpha)\mathbb{P}(f(\mathbf{x}) \neq y) \\ &+ \alpha\left|\mathbb{P}(f(\mathbf{x}) = 1|g = 1) - \mathbb{P}(f(x) = 1|g = 0)\right| \end{aligned}$$

Assuming the optimal f_F has higher positive rate for group 1 (the group 0 holds symmetrically), the fair objective function can be simplified to,

$$\begin{aligned} &(1-\alpha)\sum_{\mathbf{x} \in \mathcal{X}} \left((1 - f(\mathbf{x}))\mathbb{P}(y = 1|\mathbf{x}) + f(\mathbf{x})\mathbb{P}(y = 0|\mathbf{x})\right)\mathbb{P}(\mathbf{x}) \\ &+ \alpha \sum_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \left(\frac{\mathbb{P}(g = 1|\mathbf{x})}{\mathbb{P}(g = 1)} - \frac{\mathbb{P}(g = 0|\mathbf{x})}{\mathbb{P}(g = 0)}\right)\mathbb{P}(\mathbf{x}) \end{aligned}$$

Thus $f_F(\mathbf{x}) = 1$ is optimal if

$$\begin{aligned} &\alpha \frac{(\mathbb{P}(g = 1|\mathbf{x}) + (\mathbb{P}(g = 1) - 2)\mathbb{P}(g = 1))}{(1 - \mathbb{P}(g = 1))\mathbb{P}(g = 1)} \\ &- (1-\alpha)2\mathbb{P}(y = 1|\mathbf{x}) + 1 \geq 0 \end{aligned} \tag{1}$$

and $f_C(\mathbf{x}) = 1$ is optimal if $\mathbb{P}(y = 1|\mathbf{x}) \geq \mathbb{P}(y = 1)$. Thus, the only case in which f_F positively classifies an example $\mathbf{x}$, which is negatively classified by f_C (i.e., $f_f(\mathbf{x}) = 1 \neq f_C(\mathcal{X}) = 0$), is when the left-hand side of Inequality 1 is nonnegative and $\mathbb{P}(y = 1|\mathbf{x}) \geq \mathbb{P}(y = 1)$. Simplifying the condition in Equation **??** yields α^*. □

The key observation from Theorem 5.11 is that fairness reversal is a *small-α* phenomenon. This may seem surprising, since f_F is likely to be most similar to f_C for smaller values of α (in particular, the two are identical when $\alpha = 0$). However, when α is high, the fairness term is sufficiently dominant that reversals are unlikely. Consequently, it is precisely the intermediate values of α, where we aspire to preserve high accuracy while improving group-fairness that are most susceptible to fairness reversal. This is indeed consistent with our empirical observations in Section 4, which indicate that for intermediate values of α fairness reversals are not only more common, but occur with greater magnitude. Lastly, note that for some distributions, $\alpha^* \leq 0$, which means that fairness reversals are luckily not guaranteed.

Remark 5.12. *For some classifiers and agent distributions, fairness reversals are straightforward to prevent. We outline several of these cases in Section D.1 of the supplement.*

6 CONCLUSION

We demonstrate a fairness-reversal phenomenon, where a trained-to-be fair classifier exhibits more unfairness than the conventional accuracy-maximizing one if human agents can strategically respond to a classifier. We show that a sufficient condition for observing fairness reversal is "selectivity", that is, a group-fair classifier making fewer positive predictions than its conventional counterpart. Additionally, we demonstrated that this condition of "selectivity" also results in an accuracy reversal. The aggregate of these results indicates that when fairness is achieved through an overall decrease in positive rate (compared to the conventional classifier), strategic agent behavior can lead to a reversal of the core functionality of both models (i.e., the performance based model becomes less accurate than the fair model, and the fair model becomes less fair than the fairness-agnostic model).

We view these results not as a critique of fair-learning, but rather as a caution towards the expectation of fairness guarantees when a fair classifier sees real-world deployment. The successful deployment of fair-learning models requires the consideration of many nuanced factors, strategic agent response to model choice being on such consideration. While we have outlined several necessary and sufficient conditions regarding both classifier selectivity as well as fairness reversals, a deeper investigation into when fair classifiers may suffer from such problems in cases where the classifier s designed to anticipate strategic behavior. Mitigating fairness and accuracy reversals is an important direction for future work.

ACKNOWLEDGMENTS

This work was partially supported by the NSF (IIS- 1939677, IIS-1903207, IIS-1905558, IIS-2127752, IIS-2127754, IIS-2143895, and IIS-2040800), ARO (W911NF1810208), Amazon, and JP Morgan.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*. 60–69.
[2] Ifeoma Ajunwa, Sorelle A. Friedler, C. Scheidegger, and S. Venkatasubramanian. 2016. Hiring by Algorithm: Predicting and Preventing Disparate Impact.
[3] Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. 2020. Manipulation-proof machine learning. *arXiv preprint* (2020).
[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016), 4349–4357.
[5] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.
[6] Yiling Chen, Yang Liu, and Chara Podimata. 2020. Learning Strategy-Aware Linear Classifiers. *Advances in Neural Information Processing Systems* 33 (2020), 15265–15276.
[7] Yatong Chen, Jialu Wang, and Yang Liu. 2020. Strategic recourse in linear classification. *arXiv preprint arXiv:2011.00355* (2020).
[8] YooJung Choi, Golnoosh Farnadi, Behrouz Babaki, and Guy Van den Broeck. 2020. Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns. In *AAAI Conference on Artificial Intelligence*, Vol. 34. 10077–10084.
[9] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[10] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv preprint.
[11] Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. *Information Systems/Algoritmi* (2008).
[12] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. 2018. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 55–70.
[13] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml
[14] Cindy Durtschi, William Hillison, and Carl Pacini. 2004. The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting* 5, 1 (2004), 17–34.
[15] Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. 2021. Incentivizing Truthfulness Through Audits in Strategic Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
[16] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259–268.
[17] Stephen Giguere, Blossom Metevier, Bruno Castro da Silva, Yuriy Brun, Philip Thomas, and Scott Niekum. 2022. Fairness guarantees under demographic shift. In *International Conference on Learning Representations*.
[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.
[19] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Innovations in Theoretical Computer Science*.
[20] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016), 3315–3323.
[21] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. 2019. The disparate effects of strategic classification. In *Conference on Fairness, Accountability, and Transparency*.
[22] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. 2011. Adversarial machine learning. In *ACM Workshop on Security and Artificial Intelligence*. 43–58.
[23] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*. 1–6. https://doi.org/10.1109/IC4.2009.4909197
[24] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *International Conference on Machine Learning*. 2564–2572.
[25] Jon Kleinberg and Manish Raghavan. 2020. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)* 8, 4 (2020), 1–23.
[26] Ron Kohavi et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.. In *Kdd*, Vol. 96. 202–207.
[27] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Neural Information Processing Systems* 30 (2017).
[28] Bo Li and Yevgeniy Vorobeychik. 2018. Evasion-robust classification on binary domains. *ACM Transactions on Knowledge Discovery from Data* 12, 4 (2018), 1–32.
[29] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. 3150–3158.
[30] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. 641–647.
[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.
[32] Smitha Milli, John Miller, Anca D. Dragan, and Moritz Hardt. 2019. The social cost of strategic classification. In *Conference on Fairness, Accountability, and Transparency*. 230–239.
[33] Alan Mishler and Niccolò Dalmasso. 2022. Fair When Trained, Unfair When Deployed: Observable Fairness Measures are Unstable in Performative Prediction Settings. *arXiv preprint arXiv:2202.05049* (2022).
[34] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*. 7599–7609.
[35] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. *CoRR* abs/1709.02012 (2017). arXiv:1709.02012
[36] Michael Redmond and Alok Baveja. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141, 3 (2002), 660–678.
[37] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. 2021. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 9419–9427.
[38] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 3–13.
[39] Leanne ten Brinke, Jooa Julia Lee, and Dana R Carney. 2015. The physiology of (dis) honesty: does it impact health? *Current Opinion in Psychology* 6 (2015), 177–182.
[40] Liang Tong, Bo Li, Chen Hajaj, Chaowei Xiao, Ning Zhang, and Yevgeniy Vorobeychik. 2019. Improving robustness of ML classifiers against realizable evasion attacks using conserved features. In *USENIX Security Symposium*. 285–302.
[41] Yevgeniy Vorobeychik and Murat Kantarcioglu. 2018. *Adversarial Machine Learning*. Morgan & Claypool Publishers.
[42] L.F. Wightman and Law School Admission Council. 1998. *LSAC National Longitudinal Bar Passage Study*. Law School Admission Council. https://books.google.com/books?id=O9A7AQAAIAAJ
[43] Han Xu, Xiaorui Liu, Yaxin Li, Anil K Jain, and Jiliang Tang. 2021. To be robust or to be fair: towards fairness in adversarial training. In *International Conference on Machine Learning*.
[44] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research* 20, 75 (2019), 1–42.
[45] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *International Conference on Machine Learning*. 325–333.