"Un"Fair Machine Learning Algorithms

Runshan Fu, Manmohan Aseri, Param Vir Singh, Kannan Srinivasan
Carnegie Mellon University
{runshanf, maseri, psidhu, kannans}@andrew.cmu.edu

Machine Learning algorithms are becoming widely deployed in real world decision-making. Ensuring fairness in algorithmic decision-making is a crucial policy issue. Current legislation ensures fairness by barring algorithm designers from using demographic information in their decision-making. As a result, the algorithms need to ensure equal treatment to be legally compliant. However, in many cases, ensuring equal treatment leads to disparate impact particularly when there are differences among groups based on demographic classes. In response, several "fair" machine learning algorithms that require impact parity (e.g., equal opportunity) have recently been proposed to adjust for the societal inequalities; advocates propose changing the law to allow the use of protected class-specific decision rules. We show that these "fair" algorithms that require impact parity, while conceptually appealing, can make everyone worse off, including the very class they aim to protect. Compared to the current law, which requires treatment parity, these "fair" algorithms, which require impact parity, limit the benefits of a more accurate algorithm for a firm. As a result, profit maximizing firms could under-invest in *learning,* i.e., improving the accuracy of their machine learning algorithms. We show that the investment in learning decreases when misclassification is costly, which is exactly the case when greater accuracy is otherwise desired. Our paper highlights the importance of considering strategic behavior of stake holders when developing and evaluating "fair" machine learning algorithms. Overall, our results indicate that "fair" algorithms that require impact parity, if turned into law, may not be able to deliver some of the anticipated benefits.

Key words: Algorithmic Bias, Economics of AI, Fair ML

1. Introduction

Firms and institutions are increasingly using machine learning algorithms to make decisions in areas that have far reaching effects, such as access to credit, employment opportunities and education. Anecdotal evidence as well as recent research has highlighted concerns related to potential discrimination by algorithms (Chouldechova 2017, Fu et al. 2020, Kleinberg et al. 2016). For example, when Apple released its credit card, there were claims that women were given a lower credit limit (Washington Post 2019). Similarly, ProPublica analyzed a risk assessment software known as COMPAS that is used by judges in the United States to predict recidivism risk for an accused, and concluded that the COMPAS predictions are biased against black defendants (ProPublica 2014).

Anti-discrimination laws in the United States have been established to regulate discriminatory behavior based on protected attributes since 1964.[1] The current legislation recognizes two doctrines of discrimination: *disparate treatment* and *disparate impact*. Disparate treatment addresses procedural discrimination; it recognizes liability for treating people differently because of their membership in a protected class (e.g., race or gender) and intent to discriminate. In the algorithmic decision making context, this suggests that any explicit use of sensitive attributes, either in constructing algorithmic predictions or in setting thresholds, is strictly prohibited (Barocas and Selbst 2016). In other words, the law that prohibits disparate treatment protects individuals who are affected by algorithmic decision making against explicit discrimination.

In contrast, disparate impact addresses outcome discrimination; it recognizes liability for practices with uneven impacts on different classes (Barocas and Selbst 2016). While disparate treatment in algorithmic decision making is easy to identify, disparate impact is not. The issue is further complicated by "business necessity" as a legitimate defense to disparate impact(Chandler 1979). Hence, even though the current law has provisions for preventing disparate impact, there are little concrete guidelines on how to enforce it. Thus, the current standard practice in algorithmic decision making is to prevent disparate treatment, i.e., excluding protected attributes from inputs. This notion of fairness is known as "equal treatment" as it specifies that observationally equal individuals should be treated equally irrespective of their demographic membership (Corbett-Davies and Goel 2018).

Recently, several empirical studies have shown that enforcing equal treatment in algorithms often leads to different outcomes across demographic groups when there are systematic differences in groups (Angwin et al. 2016, Chouldechova et al. 2018, Fuster et al. 2017, Skeem and Lowenkamp 2016). In response, several fairness notions, such as equal opportunity, demographic parity, equalized odds, and conditional statistical parity, have been proposed with the aim to ensure certain perspectives of impact parity in algorithmic decision-making. The corrections for disparate impact usually require treating different groups differently, thus violating equal treatment. As a result a policy debate has ensured as to whether algorithms should be required to satisfy "treatment parity" or "impact parity" (Barocas and Selbst 2016, Corbett-Davies and Goel 2018, Hardt et al. 2016, Skeem and Lowenkamp 2016, Kim 2017).

Impact parity can be violated from many different dimensions. Thus, there is no single fairness notion that captures the absolute impact parity. In fact, Kleinberg et al. (2016) and Chouldechova (2017) show that some popular fairness notions that focus on impact parity cannot be satisfied simultaneously except in highly constrained special cases. However, one fairness notion, *equal opportunity*,

[1] Civil Rights Act of 1964 (Pub. L. 88-352) (Title VII) is generally viewed to be the first major development in anti-discrimination law in the US.

has received considerable attention (Hardt et al. 2016).[2] Equal opportunity requires parity in the proportion of positive decisions in deserving individuals (e.g., loan approval among non-defaulters).[3] The underlying idea is that qualified individuals should be given equal opportunity to access a desirable outcome, regardless of their demographic attributes. Equal opportunity is appealing, as it allows the protected group, which has been historically discriminated against, to have equal access to opportunities. In this paper, we focus on equal opportunity as a representative fair condition for the notion of equal impact.

Our main research goal is to investigate whether regular and protected groups are better off if the legislation were to change to require equal opportunity instead of equal treatment for algorithmic decision-making. Further, what would be the firm's optimal learning efforts under the two different fairness notions? Finally, would the decision-maker be better off if equal opportunity was required instead of equal treatment?

The highlight of this paper is the consideration of strategic incentives of the decision-maker. When comparing the effect of equal treatment to equal impact on the regular and protected groups, extant research generally takes the trained machine learning algorithm as given and focus on making decisions that satisfy a particular notion of fairness. However, in reality, the accuracy of algorithms depends on the amount of learning effort that decision-makers (firms or institutions) exert. To learn more accurately about the outcome of interest, firms and institutions need to collect high quality and relevant data, build and improve infrastructure, experiment, develop and update machine learning models, etc. In return, more accurate predictions allow them to make better decisions that increase their utilities. Thus, firms and institutions choose the optimal amount of learning effort that maximizes their profits. To our knowledge, existing research has not studied how decision-makers would respond to a policy that requires equal treatment compared to one that requires equal impact. We address this important gap in literature and compare the effect of the two fairness notions in the case where the firm endogenously chooses an optimal level of learning effort.

We present a parsimonious theoretical model where a risk-neutral decision-maker wants to select good candidates. A candidate is either good or bad. Further, the candidate either belongs to a regular or a protected group. The accuracy of the machine learning model depends upon the learning cost that the decision-maker (i.e., firm) incurs. A higher investment (learning cost) in the algorithm leads to greater accuracy. The two notions of fairness - equal treatment and equal opportunity - are enforced as separate constraints that the decision-maker must satisfy. We solve for optimal learning

[2] Some of the other fairness notions are Demographic parity, Equalized odds and Predictive rate parity

[3] The notion of equal opportunity violates Title VII of the United States Civil Rights Act. This is one of the main reasons that equal opportunity is just a proposal, not an actual law.

effort, firm profit and approval rates for protected and regular groups under the equal treatment and equal opportunity regimes. We first carry out the comparison between equal treatment and equal opportunity. Then we compare equal treatment to a general fairness notion that simply requires a lower threshold (more approval) for the protected group.

The key difference between the regular group and the protected group in our model is that the algorithm can better separate good applicants from bad applicants for the regular group than for the protected group. In other words, the algorithm's learning efficiency is higher for the regular group than the protected group. Several empirical analyses have shown that machine learning algorithms usually learn more efficiently for the regular group than the protected group, meaning that the predictions from the same algorithm tend to be more accurate for the regular group (Hardt et al. 2016, Chouldechova and G'Sell 2017). Therefore, the risk distributions for good and bad applicants are usually less overlapped in the regular group, reflecting a better separation. There are several reasons for this unequal separation:

- The same features in the regular group and the protected group may have different relationships to the outcome of interest, and machine learning models may not be flexible enough to capture the difference, especially when they are unaware of the group membership (Chouldechova and Roth 2018, Hardt 2014).
- Features that better account for pertinent statistical variation among members of the protected class are usually more expensive to collect and, hence, ignored. Further, the quality of data records is usually lower for the protected group (Barocas and Selbst 2016).
- The protected group tends to be underrepresented in the data (Chouldechova and Roth 2018, Barocas and Selbst 2016, Hardt 2014). First, the protected group may account for a smaller portion in the population. For example, blacks are the protected compared to whites. Second, the protected group has disproportionately less presence in the data due to historically biased decisions. For example, if females have had lower chances of getting loans, they would appear less frequently than males in the data.

1.1. Equal Treatment versus Equal Impact

Before diving into our main results, let us look at an example to see how satisfying equal treatment would lead to disparate impact when a machine learning algorithm learns more efficiently about the regular group than the protected group. Consider a bank making loan-granting decisions. There are two classes of loan applicants: a regular and a protected class (e.g., male and female), and each class contains 10 applicants. Some applicants will pay back the loan (non-defaulters) and others will not (defaulters). The bank wants to give loans to the non-defaulters. However, defaulters and non-defaulters are not easily identifiable. To aid its decisions, the bank collects data, builds an algorithm

that estimates the probability of payback for the 20 applicants, and gives loans to the applicants with high probability of payback. The bank is unbiased and abides by the law of equal treatment – it does not use the class membership in the prediction algorithm and it sets the same thresholds for both groups.

An ideal scenario is that the algorithm perfectly separates defaulters and non-defaulters. Unfortunately, perfect separation is impossible in real life. Currently, state-of-art machine learning algorithms can only achieve a 70% - 80% accuracy rate for tasks such as loan default or criminal recidivism predictions (Kleinberg et al. 2017, Netzer et al. 2018). Another ideal scenario is that the risk distribution is independent of the class attribute. In this case, the probability distributions for the regular group and the protected group are identical, as illustrated in figure 1. Setting a single threshold, in this case, at any value for both groups is fair from virtually all perspectives. Unfortunately, this case is rare if not impossible.

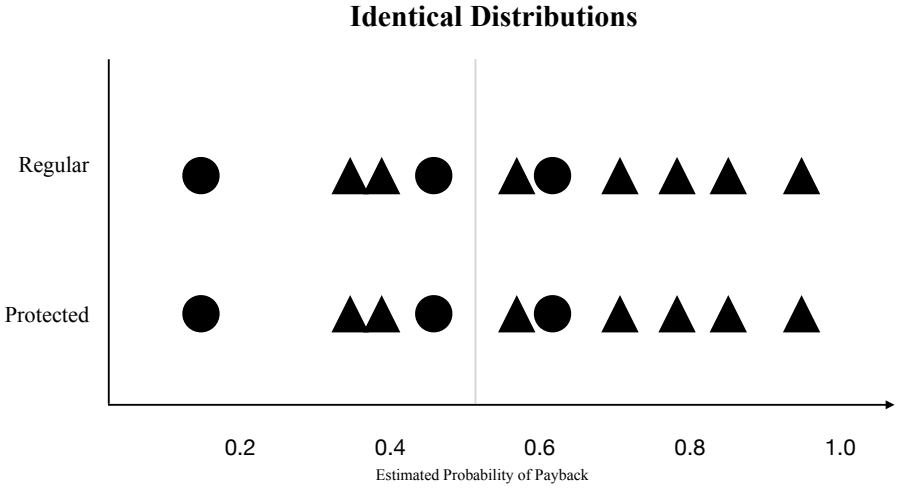

Figure 1 **The estimated probability of payback for the 20 applicants when the risk is independent of class membership. Each circle represents a defaulter and each triangle represents a non-defaulter. The regular group and the protected group have identical probability distribution, just as two random samples.**

Therefore, the realistic scenario is that the algorithm cannot perfectly separate the defaulters and non-defaulters, and it produces different risk distributions for the regular group and the protected group, as depicted in Figure 2. Now, if the bank uses the profit maximizing single threshold for both groups (say 0.5), 60% of the regular applicants will get the loan, while only 50% of the protected applicants will be approved. Moreover, among the qualified applicants – the non-defaulters – the chance of getting the loan is about 71.4% in the regular group, but only 50% in the protected group. While the bank's choice of single threshold appears neutral to the two groups, it has a disadvantageous effect on the protected group leading to disparate impact.

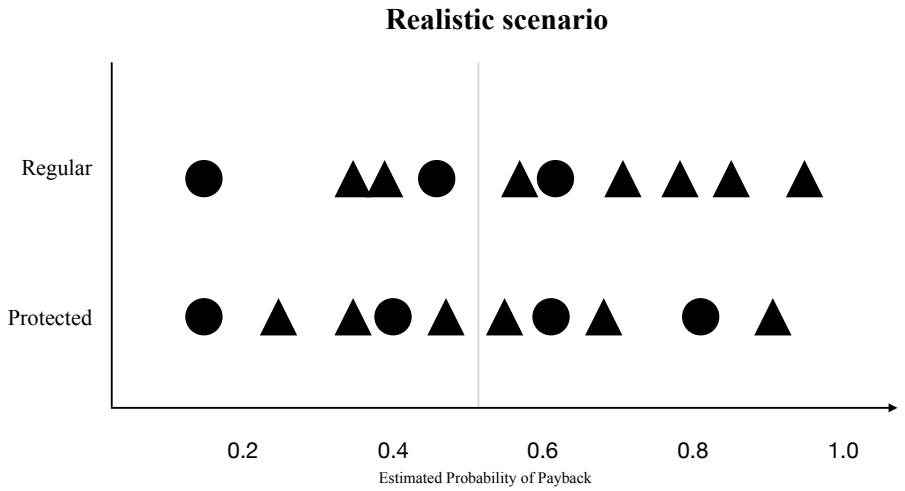

Figure 2 **The estimated probability of payback for the 20 applicants in a realistic scenario. Each circle represents a defaulter and each triangle represents a non-defaulter.**

1.2. Main Results Synopsis

Our first result is that **the firm's optimal learning effort is lower under equal opportunity as compared to that under equal treatment.** In other words, the firm's algorithm will be less accurate under equal opportunity than under equal treatment. The key intuition behind this result is as follows: At any learning effort (i.e., given a machine learning algorithm), the good and bad candidates are less separated for protected group compared to the regular group. As a result, the protected group represents a riskier pool compared to the regular group. Hence, given equal learning effort, compared to equal treatment, equal opportunity requires the firm to remove candidates from a less risky regular pool and/or accept candidates from a more risky protected pool. As a result, the firm is able to extract greater returns from learning effort and therefore is incentivized to learn more under equal treatment compared to under equal opportunity.

Most research on fair algorithms argues that requiring equal impact instead of equal treatment will help the protected group at the expense of the regular group (Hardt et al. 2016, Chouldechova and G'Sell 2017). However, our second result shows that **compared to equal treatment, equal opportunity can make everyone worse-off, including the protected group that it aims to help.** As discussed earlier, the proponents of fair algorithms that focus on equal impact generally ignore the firm as a strategic player. We show that when the firm strategically chooses the learning effort to maximize its profit, it could make everyone worse off. There are two effects at play here. First, as discussed with the intuition behind our first result, the firm has to let go of the less risky candidates from the regular group and accept candidates from the riskier protected group. The regular group gets hurt in this case because the threshold for them to be accepted is raised, while the

protected group benefits as the threshold for them to be accepted is lowered. Second, the optimal algorithm for the firm is less accurate under equal opportunity than equal treatment. In this case, the firm acts conservatively and raises the thresholds for both the regular and the protected group under equal opportunity, which hurts both groups by accepting fewer candidates. When the market is risky, under equal opportunity, the threshold for the protected group is raised to a higher level compared to what it was under equal treatment. As a result, while the protected group does receive equal opportunity as the regular group, both receive less opportunity under equal opportunity compared to under equal treatment.

While the trade-offs between fairness and accuracy for a decision-maker is well established, it is not clear how the decision-maker would be affected by two different fairness notions. Our third result, shows that **the firm profit would be lower under equal opportunity compared to equal treatment**. The intuition for this result is as follows: Compared to equal treatment, the firm can take less advantage of learning under equal opportunity, which lowers its profits. Moreover, the firm invests less in learning effort under equal opportunity and as a result, is not able to separate the good and bad applicants well. Hence, both the regular and the protected group pools are riskier for the firm under equal opportunity, which further hurts its profits.

There are many widely viewed benefits of equal impact that we do not capture in our main model, but address through extensions of our model. First, one argument is that the algorithm accuracy for the protected group could increase when more applicants from the protected group are approved, and equal impact may better improve fairness in the long term by reducing the learning efficiency gap over time. We extend our main model to capture the effect of approval rates for a group on their future learning efficiency. Our results show that the learning efficiency for the protected group would improve at a slower rate under equal impact than equal treatment, and our main results still hold in the extended model. The intuition behind this result directly follows from our second result. In any period, fewer members of the protected and regular group are likely to be approved under equal impact than under equal treatment. As a result, the effect on the next period learning efficiency is smaller under equal opportunity than under equal treatment. Second, the protected group members may feel that they have a better chance under equal impact and hence would improve their quality. That is, a greater fraction of protected group individuals would become good under equal impact than under equal treatment. While we do not fully model the process through which individuals invest in themselves to become good or bad candidates, for comparison purposes we do the following: We compare the case where the protected and the regular groups have the same fraction of good and bad applicants under equal opportunity to the case where the protected group has a smaller fraction of good applicants than that in the regular group under equal treatment. We show that even in this comparison, the optimal learning effort would be lower under equal impact than under equal treatment.

1.3. Contributions

Our paper makes several contributions. To our knowledge, we are the first to provide a framework to compare the effect of two countervailing fairness notions on the decision-maker and the individuals that are affected by these decisions when accounting for the strategic role of the decision-maker in algorithmic decision-making. We are also one of the first to endogenize the accuracy of the model. Extant literature has ignored the strategic role that a decision maker plays and the cost of learning in this context. Modeling the incentives of the decision-maker helps us revert some of the results that are commonly accepted in the fair machine learning literature. By definition, the equal opportunity or other related fair algorithms that enforce equal impact, help the protected group at the cost of regular group. However, we show that when the market is risky, both the regular and the protected group could be worse off under equal opportunity compared to under equal treatment, because equal opportunity discourages learning effort.

As the strategic role of the decision-maker is ignored in the fair machine learning literature, the impact of a specific fairness notion on the decision-maker is also typically not considered. The fairness-efficiency trade-off is widely accepted. Thus, it is accepted that enforcing fairness of any form could make the decision-maker less efficient. However, how different fairness notions would affect the profit of the decision maker is not that obvious. We show that the decision-maker would be worse under equal opportunity compared to equal treatment.

Our results are not limited to the notion of "equal opportunity" only. Any fairness notion that closes the gap of learning outcomes by lowering the threshold for the protected group (relative to the regular group) would reduce the firm's learning effort, and therefore harm the firm, the regular group and sometimes even the protected group too. Our results highlight that these different "fair" algorithms that aim to ensure equal impact, if turned into law in place of equal treatment, may not be able to deliver some of the anticipated benefits.

In a number of extensions of our main model, we show that our results are robust even when considering several widely viewed benefits of equal impact over equal treatment. One may think that the difference in learning efficiencies of the regular and the protected group is a short run observation, and in the long run, the difference will disappear as more data becomes available. Yet, we show that when the algorithmic accuracy is a function of past approval rates, it will improve at a slower rate under fairness notions motivated by equal impact than under equal treatment. Another argument in support of fairness notions motivated by equal impact is that the protected group would feel they have a better chance under these conditions and as a result invest in improving their underlying distribution. In other words, the distribution of good and bad applicants in the protected group may improve to match that of the regular group under these fairness constraints. In an extension of

our main model, we show that even when the protected group improves its distribution under equal opportunity, the optimal learning effort of the decision maker is still lower compared to that under equal treatment.

The next section presents our model. Sections 3 and 4 analyze and compare the decision maker's optimal decisions under equal treatment and under equal opportunity. In Section 5, we model a general fairness notion with the property that it lowers the threshold for the protected group. We show that our results continue to hold for this fairness notion too. Section 6 presents robustness check under additional features. Section 7 concludes.

2. Model

Consider a decision-maker who needs to make the decision of accepting or not accepting a candidate (e.g., a bank deciding to approve the loan for an applicant, a university deciding to admit a student). A candidate can be *good* or *bad*. For the decision-maker, the utility of accepting a good candidate is α, and the dis-utility of accepting a bad candidate is β. If the decision-maker does not accept the candidate, it will earn zero utility from that candidate. The decision maker can exert effort (invest in learning) to separate good candidates from bad candidates. This is different from the model of Shimao et al. (2018), which models the effort exerted by applicants in improving their quality. We use s to denote such learning effort, and use $\tau(s)$ for the cost of exerting an effort of s. We assume that $\tau(s)$ is a convex increasing function of s, which means that it is increasingly costly to better learn and separate good candidates from bad ones.

Based on a sensitive attribute (such as race or gender), each candidate belongs to one of two groups: a protected group and a regular group. The protected group is the group of candidates that needs to be *protected* against any potential discrimination. All other candidates are in the regular group. We assume that both groups are of equal size. The number of regular, as well as protected, candidates is normalized to 1. Both groups contain good and bad candidates. Let d_p and d_r be the proportion of bad candidates for the protected and the regular group, respectively. We assume that $d_p \geq d_r$. The underlying reason is that the protected group is usually at a disadvantage in terms of socioeconomic status (due to complex historical reasons), which makes it more difficult for a protected candidate to be a good candidate (by paying back the loan, for example). Note that this assumption includes the case where the two groups have the same proportion of bad candidates, i.e., $d_p = d_r$. Define $\alpha_p = \alpha(1 - d_p)$, $\beta_p = \beta d_p$, and similarly $\alpha_r = \alpha(1 - d_r)$, $\beta_r = \beta d_r$. We assume that $\beta_r \geq \alpha_r$, and because $d_p \geq d_r$, this implies that $\beta_p \geq \alpha_p$. Intuitively, this assumption means that if the decision-maker accepts all candidates in either group, then it derives a negative expected utility ($\alpha_r - \beta_r \leq 0$ and $\alpha_p - \beta_p \leq 0$). Compared to Shimao et al. (2019), who analyze the equilibrium

behavior of algorithms when the subjects (e.g., applicants) are strategic, we do not model agents as strategic. That is, the agents cannot change their type (good or bad). In section 6.2, we do a robustness check to show that our main results still hold when we consider the type change caused by applicants' strategic response.

The decision-maker assigns scores to the candidates to represent their goodness. The distributions of scores depend on the learning effort s: the higher s is, the better separated are bad candidates and good candidates. Figure 3 pictorially depicts score distributions for different groups. [4] We assume that the scores of protected-bad candidates and protected-good candidates are uniformly distributed as $U[0,\ 1-\gamma_p s]$ and $U[\gamma_p s,\ 1]$ respectively. Intuitively, when the decision maker does not exert any effort to separate bad candidates from good candidates, i.e., when $s = 0$, the scores of both bad candidates and good candidates are uniformly distributed between 0 and 1. As the decision-maker starts exerting effort (i.e., $s > 0$) to learn and separate the two sub-groups, the bad candidates and good candidates start to separate. The rate of this separation is captured by the parameter $\gamma_p > 0$ (the subscript p stands for the protected group). Similarly, the scores of regular-bad candidates and regular-good candidates are uniformly distributed as $U[0,\ 1-\gamma_r s]$ and $U[\gamma_r s,\ 1]$ respectively, where γ_r represents the rate of separation of bad candidates and good candidates in the regular group. Because the learning efficiency is higher for the regular group, we have $\gamma_r > \gamma_p$. To avoid trivial cases, we assume perfect separation would never happen for either group, i.e., the score distributions for bad candidates and good candidates always have overlaps. Mathematically, this means the maximum amount of learning is less than $\frac{1}{2\gamma_r}$, i.e. $s \in [0, \frac{1}{2\gamma_r})$.

Empirical Evidence for the Assumption $\gamma_r > \gamma_p$: We test this assumption on a real-world dataset from Prosper, which is a crowdfunding platform. The data set contains 3776 observations. Each observation represents a loan application. We tried to predict whether an applicant pays back the loan or not, using several attributes such as Listing Amount, Credit Score, Debt, Income, etc. The prediction was performed using several state-of-the-art algorithms such as XGBoost, SVM, Random Forest, Naive Bayes, and Multi-Layer Perceptions. The best performing algorithm was XGBoost, and its accuracy for males was 0.7344, and for females, it was 0.6479. Thus, there is a gap of 0.0865 between the accuracies for males and females. Similar gaps exist for other algorithms too. For example, the XGBoost accuracy for males was 0.7905 and for females was 0.7375 for predicting credit risk on a publicly available German Credit Data at the UCI Machine learning repository. In another publicly available dataset from the UCI repository that captures student achievement in secondary education of two Portuguese schools, XGBoost results report a mean squared error of 3.1982 for females and 1.6451 for males when predicting their math grades.

[4] The height of the bars corresponding to "good" and "bad" applicants is only for visualization; it does not necessarily mean that there are more "bad" applicants. This height depends on the default rate and other parameters of the model.

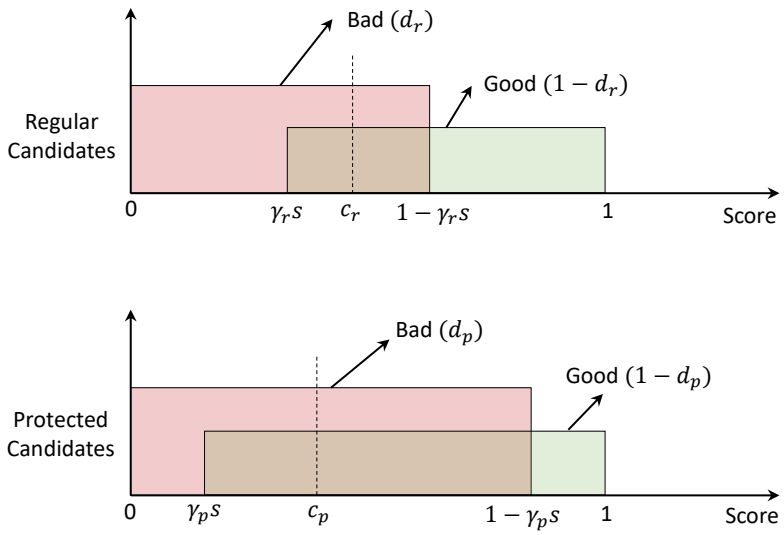

Figure 3 **Distribution of scores of bad candidates and good candidates in protected and regular groups. The same value of s achieves more separation of good and bad candidates in the regular group compared to the protected group.**

The decision-maker employs *threshold rules* when accepting the candidates. That is, it accepts a candidate if and only if his or her score is higher than a fixed threshold. The decision-maker decides two thresholds, c_p and c_r, for the protected group and the regular group, respectively. The threshold rule is intuitive and widely used in practice, and Corbett-Davies and Goel (2018) demonstrates that the threshold rule based on true risk produces the optimal decisions for a rational decision maker.

The two thresholds, c_p and c_r, along with the learning effort, s, are chosen jointly to maximize the decision-maker's profit. In the absence of the fairness requirement, the decision-maker can choose the three decisions freely within their feasible ranges. When certain fairness notions are enforced, the decision-maker has to choose the decisions in a way that satisfies the fairness requirement. Mathematically, each fairness definition is a constraint in the decision-maker's optimization problem, and the decision-maker's optimal decisions (i.e., c_p, c_r, and s) will change according to the fairness constraint that it operates under.

- **Equal Treatment (ET):** Use the same threshold for both the groups, i.e.,

$$c_p = c_r.$$

  Under this constraint, the sensitive group label is not used in the decision. All candidates are treated equally, as they are held to the same standard, irrespective of their group identity. Equal Treatment is currently required by law. Failure to satisfy it would constitute a violation of treatment parity.

- **Equal Opportunity (EO):** Given that a candidate is a good candidate, the probability of getting accepted should be the same for the protected and the regular candidates Hardt et al. (2016), i.e.,

$$\text{Prob}_p[L=1|D=0] \;=\; \text{Prob}_r[L=1|D=0],$$

  where $L=1$ represent that the candidate is accepted, $D=0$ represents that the candidate is a good, and $\text{Prob}_p[\cdot]$ and $\text{Prob}_r[\cdot]$ represents the probabilities corresponding to protected and regular groups respectively. This fairness notion requires that the truly "qualified" candidates (good candidates) have an equal probability of getting accepted in both the groups.
- **Demographic Parity (DP):** The probability of getting accepted should be the same for protected and regular candidates, i.e.,

$$\text{Prob}_p[L=1] \;=\; \text{Prob}_r[L=1].$$

Since the notion of Equal Opportunity (EO) and Demographic Parity (DP) are both group-level fairness notions, they are qualitatively very similar. Thus, we only consider the first two fairness notions for our analysis, i.e., Equal Treatment and Equal Opportunity, as representative fairness-notions of individual and group fairness, respectively. In a later section, we compare equal treatment with a general equal impact fairness algorithm.

The first central question we address in this paper is which of these fairness notions leads to more learning, where the amount of learning is quantified as the optimal value of s that the decision-maker chooses. The amount of learning determines the distributions of the scores, and therefore also influences the thresholds, c_p and c_r and the final acceptance decisions. Hence, the second central question we address is – how do these fairness notions affect the decision maker's profit and the opportunity of getting accepted for the candidates.

3. Analysis

We now proceed with analyzing the decision-maker's optimization problem under the two fairness constraints. The decision-maker chooses three quantities, s, c_p and c_r, to maximize its profit. The decision-maker itself is unbiased, in the sense that it only cares about the profit, and candidates in the protected and the regular groups contribute to the profit in the same way conditional on whether they are good or bad candidates. Let π represent the profit of the decision-maker, then we have

$$\begin{aligned}\pi \;=\; & \min\left\{\frac{1-c_r}{1-\gamma_r s},\ 1\right\}\cdot\alpha_r - \max\left\{\frac{1-\gamma_r s-c_r}{1-\gamma_r s},\ 0\right\}\cdot\beta_r \\ & +\ \min\left\{\frac{1-c_p}{1-\gamma_p s},\ 1\right\}\cdot\alpha_p - \max\left\{\frac{1-\gamma_p s-c_p}{1-\gamma_p s}, 0\right\}\cdot\beta_p - \tau(s) \qquad (1)\end{aligned}$$

Table 1 The main notation for our analysis

Notation	Description
α	The utility of selecting a good candidate.
β	The dis-utility of selecting a bad candidate.
$d_p,\ d_r$	Fraction of bad candidates in the protected and regular groups.
$\alpha_p,\ \alpha_r$	Expected utility of selecting a good candidate in protected and regular groups, respectively: $\alpha_p = \alpha(1-d_p),\ \alpha_r = \alpha(1-d_r)$.
$\beta_p,\ \beta_r$	Expected dis-utility of selecting a bad candidate in protected and regular groups, respectively: $\beta_p = \beta d_p,\ \beta_r = \beta d_r$.
c_r	Decision threshold for the regular group.
c_p	Decision threshold for the protected group.
s	Effort exerted by the decision-maker.
$\tau(s)$	Cost of exerting an effort of s.
γ_p	Rate of learning for the protected group.
γ_r	Rate of learning for the regular group.

For each fairness constraint, we first obtain the optimal value of c_p and c_r for a given s, and then write the profit of decision-maker as a function of s. These functions give the highest possible profit (achieved by choosing the optimal thresholds) for any value of learning effort s. Intuitively, more learning (higher s) means better separation of good and bad candidates, and therefore higher revenue, but it also leads to higher cost since $\tau(s)$ is an increasing function of s. Next, we analyze the decision maker's problem under the fairness notion of equal-treatment.

3.1. Equal Treatment

In this case, the decision-maker maximizes its profit given in Equation (1) such that $c_p = c_r$. We use c to denote the common threshold for both the groups, where $c = c_p = c_r$. Let c^{ET} represent the optimal common threshold. Note that $c^{\mathrm{ET}} \in [\gamma_p s, 1-\gamma_p s]$, because when the common threshold is lower than $\gamma_p s$ (or higher than $1-\gamma_p s$), the decision-maker can always increase (or decrease) it to achieve a higher profit.

Define the following:

When $c \in [\gamma_p s, 1-\gamma_p s]$, we can rewrite the profit function under equal treatment constraint as:

$$\pi_{\mathrm{ET}} = \begin{cases} \left(\frac{\beta_r}{1-\gamma_r s} + \frac{\beta_p-\alpha_p}{1-\gamma_p s}\right) c + \frac{\alpha_p}{1-\gamma_p s} + \alpha_r - \beta_r - \beta_p - \tau(s), & \text{if } \gamma_p s \le c < \gamma_r s, \\ \left(\frac{\beta_r-\alpha_r}{1-\gamma_r s} + \frac{\beta_p-\alpha_p}{1-\gamma_p s}\right) c + \frac{\alpha_r}{1-\gamma_r s} + \frac{\alpha_p}{1-\gamma_p s} - \beta_r - \beta_p - \tau(s), & \text{if } \gamma_r s \le c < 1-\gamma_r s, \\ \left(\frac{\beta_p-\alpha_p}{1-\gamma_p s} - \frac{\alpha_r}{1-\gamma_r s}\right) c + \frac{\alpha_r}{1-\gamma_r s} + \frac{\alpha_p}{1-\gamma_p s} - \beta_p - \tau(s), & \text{if } 1-\gamma_r s \le c \le 1-\gamma_p s. \end{cases} \tag{2}$$

The derivative with respective to c is

$$\frac{d\pi_{\text{ET}}}{dc} = \begin{cases} \frac{\beta_r}{1-\gamma_r s} + \frac{\beta_p - \alpha_p}{1-\gamma_p s}, & \text{if } \gamma_p s \le c < \gamma_r s, \\ \frac{\beta_r - \alpha_r}{1-\gamma_r s} + \frac{\beta_p - \alpha_p}{1-\gamma_p s}, & \text{if } \gamma_r s \le c \le 1-\gamma_r s, \\ \frac{\beta_p - \alpha_p}{1-\gamma_p s} - \frac{\alpha_r}{1-\gamma_r s}, & \text{if } 1-\gamma_r s \le c \le 1-\gamma_p s. \end{cases} \tag{3}$$

The profit function is piece-wise linear in c. In the first two cases of the above equation, both $\left(\frac{\beta_r}{1-\gamma_r s} + \frac{\beta_p - \alpha_p}{1-\gamma_p s}\right)$ and $\left(\frac{\beta_r - \alpha_r}{1-\gamma_r s} + \frac{\beta_p - \alpha_p}{1-\gamma_p s}\right)$ are positive because $\beta_p > \alpha_p$ and $\beta_r > \alpha_r$. Therefore, the profit is increasing in c in these two cases, i.e., in the range $[\gamma_p s, 1-\gamma_r s]$. Depending on the the sign of $\left(\frac{\beta_p - \alpha_p}{1-\gamma_p s} - \frac{\alpha_r}{1-\gamma_r s}\right)$, the optimal value of c could be either $1-\gamma_r s$ or $1-\gamma_p s$.[5] Thus, we have

$$c^{\text{ET}} = \begin{cases} 1-\gamma_r s & \text{if } \beta_p \le \alpha_p + \frac{1-\gamma_p s}{1-\gamma_r s}\alpha_r, \\ 1-\gamma_p s & \text{otherwise.} \end{cases} \tag{4}$$

Let $\pi_{\text{ET}}(s)$ be the optimal profit of the decision maker under equal treatment, for a given value of s. Substituting the optimal threshold in Equation (2), we get

$$\pi_{\text{ET}}(s) = \begin{cases} \frac{\alpha_p \gamma_r s}{1-\gamma_p s} + \frac{\alpha_r \gamma_r s}{1-\gamma_r s} + \frac{\beta_p \gamma_p s}{1-\gamma_p s} - \frac{\beta_p \gamma_r s}{1-\gamma_p s} - \tau(s), & \text{if } \beta_p < \alpha_p + \frac{1-\gamma_p s}{1-\gamma_r s}\alpha_r, \\ \frac{\alpha_p \gamma_p s}{1-\gamma_p s} + \frac{\alpha_r \gamma_p s}{1-\gamma_r s} - \tau(s), & \text{otherwise.} \end{cases} \tag{5}$$

Next, we analyze the decision-maker's problem under the fairness notion of equal opportunity.

3.2. Equal Opportunity

The equal opportunity constraint requires the same rate of acceptance for the good candidates in the protected and the regular groups. Mathematically, the constraint of equal opportunity can be written as follows:

$$\frac{1-c_r}{1-\gamma_r s} = \frac{1-c_p}{1-\gamma_p s}. \tag{6}$$

Thus, the decision maker maximizes its profit subject to the above constraint. Let c_p^{EO} and c_r^{EO} be the optimal acceptance thresholds under equal opportunity. First, we note the following about the equal opportunity constraint in (6).

Proposition 1 *Under equal opportunity, the acceptance threshold is lower for the protected group compared to the regular group, i.e.,* $c_p^{EO} < c_r^{EO}$.

[5] When $\frac{\beta_p - \alpha_p}{1-\gamma_p s} - \frac{\alpha_r}{1-\gamma_r s} = 0$, the profit is a constant on $[1-\gamma_r s, 1-\gamma_p s]$. The optimal c can be any value in the range. We take $1-\gamma_r s$ as the optimal value in this case.

Proof: Because $\gamma_p < \gamma_r$, $1-\gamma_r s < 1-\gamma_p s$. Combine the inequality with the constraint in (6), we have $1-c_r^{\text{EO}} < 1-c_p^{\text{EO}}$, which means $c_p^{\text{EO}} < c_r^{\text{EO}}$. ■

Intuitively, the separation of good and bad candidates is faster for the regular group compared to the protected group. Therefore, if the decision-maker uses the same acceptance threshold for both regular and protected candidates, then more good candidates of the regular group will get accepted compared to the protected group. However, the constraint of equal opportunity mandates that the probability of getting accepted should be the same for the good candidates of both the groups; thus, the decision-maker has to lower the threshold for the protected group to accept more protected-good candidates. In other words, for a fixed level of learning effort s, equal opportunity is in favor of the protected group.

Re-arrange the equal opportunity constraint in (6), we have

$$c_r = \frac{1-\gamma_r s}{1-\gamma_p s} c_p + \frac{(\gamma_r - \gamma_p)s}{1-\gamma_p s} \tag{7}$$

The constraint poses a one-to-one mapping between c_p and c_r: one is determined when the other one is chosen. Therefore, the decision-maker is effectively setting one threshold. In the following text, we will use c_p as the decision variable, and c_r can always be obtained using (7). Similar to Section 3.1, we note that $c_p^{\text{EO}} \in [\gamma_p s, 1-\gamma_p s]$. Substituting the value of c_r from (7) into the profit function (1), we have

$$\pi_{\text{EO}} = \begin{cases} \frac{(\beta_p-\alpha_p)+(\beta_r-\alpha_r)}{1-\gamma_p s} c_p + \frac{\alpha_p+\alpha_r-\beta_r}{1-\gamma_p s} + \frac{\beta_r}{1-\gamma_r s} - \beta_p - \beta_r - \tau(s), & \text{if } \gamma_p s \le c_p \le 1 - \frac{1-\gamma_p s}{1-\gamma_r s}\gamma_r s, \\ \frac{\beta_p-\alpha_p-\alpha_r}{1-\gamma_p s} c_p + \frac{\alpha_p+\alpha_r}{1-\gamma_p s} - \beta_p - \tau(s), & \text{if } 1 - \frac{1-\gamma_p s}{1-\gamma_r s}\gamma_r s < c_p \le 1-\gamma_p s. \end{cases} \tag{8}$$

The derivative of the profit with respect to c_p is

$$\frac{d\pi_{\text{EO}}}{dc_p} = \begin{cases} \frac{(\beta_p-\alpha_p)+(\beta_r-\alpha_r)}{1-\gamma_p s}, & \text{if } \gamma_p s \le c_p \le 1 - \frac{1-\gamma_p s}{1-\gamma_r s}\gamma_r s, \\ \frac{\beta_p-\alpha_p-\alpha_r}{1-\gamma_p s}, & \text{if } 1 - \frac{1-\gamma_p s}{1-\gamma_r s}\gamma_r s < c_p \le 1-\gamma_p s. \end{cases} \tag{9}$$

Similar to the case under equal treatment, $\frac{(\beta_p-\alpha_p)+(\beta_r-\alpha_r)}{1-\gamma_p s}$ is positive, and the optimal value of c_p depends on the sign of $\frac{\beta_p-\alpha_p-\alpha_r}{1-\gamma_p s}$. Thus, we have

$$\left(c_p^{\text{EO}}, c_r^{\text{EO}}\right) = \begin{cases} \left(1 - \frac{\gamma_r s(1-\gamma_p s)}{1-\gamma_r s},\ 1-\gamma_r s\right) & \text{if } \beta_p \le \alpha_p + \alpha_r, \\ \left(1-\gamma_p s,\ 1 - \frac{1-\gamma_r s}{1-\gamma_p s}\gamma_p s\right), & \text{otherwise.} \end{cases} \tag{10}$$

Let $\pi_{\text{EO}}(s)$ be the optimal profit of the decision-maker under equal opportunity, for a given value of s. Substituting the optimal thresholds into Equation (8), we get

$$\pi_{\text{EO}}(s) = \begin{cases} \frac{\alpha_p\gamma_r s}{1-\gamma_r s} + \frac{\alpha_r\gamma_r s}{1-\gamma_r s} + \frac{\beta_p\gamma_p s}{1-\gamma_p s} - \frac{\beta_p\gamma_r s}{1-\gamma_r s} - \tau(s), & \text{if } \beta_p \le \alpha_p + \alpha_r, \\ \frac{\alpha_p\gamma_p s}{1-\gamma_p s} + \frac{\alpha_r\gamma_p s}{1-\gamma_p s} - \tau(s), & \text{otherwise.} \end{cases} \tag{11}$$

From the above analysis, we can see that, both under equal treatment and under equal opportunity, the optimal threshold and the resulting profit function in learning-effort depend on β_p, the expected loss from the bad candidates in the protected group. Intuitively, when this expected loss is too high (either because the loss due to selecting a bad candidate β is large compared to the benefit of selecting a good candidate α, or the fraction of bad candidates d_p is high), the decision maker would be conservative and pick up a higher threshold. Otherwise, the decision maker would choose a lower threshold to accept more candidates. Specifically, we can summarize the results into three cases:

(1) When $\beta_p \leq \alpha_p + \alpha_r$,

$$c_p^{\mathrm{ET}} = c_r^{\mathrm{ET}} = 1 - \gamma_r s, \qquad \pi_{\mathrm{ET}}(s) = \frac{\alpha_p\gamma_r s}{1-\gamma_p s} + \frac{\alpha_r\gamma_r s}{1-\gamma_r s} + \frac{\beta_p\gamma_p s}{1-\gamma_p s} - \frac{\beta_p\gamma_r s}{1-\gamma_p s} - \tau(s)$$

$$c_p^{\mathrm{EO}} = 1 - \frac{\gamma_r s(1-\gamma_p s)}{1-\gamma_r s},\ c_r^{\mathrm{EO}} = 1-\gamma_r s, \quad \pi_{\mathrm{EO}}(s) = \frac{\alpha_p\gamma_r s}{1-\gamma_r s} + \frac{\alpha_r\gamma_r s}{1-\gamma_r s} + \frac{\beta_p\gamma_p s}{1-\gamma_p s} - \frac{\beta_p\gamma_r s}{1-\gamma_r s} - \tau(s);$$

(2) When $\alpha_p + \alpha_r < \beta_p \leq \alpha_p + \frac{1-\gamma_p s}{1-\gamma_r s}\alpha_r$,

$$c_p^{\mathrm{ET}} = c_r^{\mathrm{ET}} = 1 - \gamma_r s, \qquad \pi_{\mathrm{ET}}(s) = \frac{\alpha_p\gamma_r s}{1-\gamma_p s} + \frac{\alpha_r\gamma_r s}{1-\gamma_r s} + \frac{\beta_p\gamma_p s}{1-\gamma_p s} - \frac{\beta_p\gamma_r s}{1-\gamma_p s} - \tau(s)$$

$$c_p^{\mathrm{EO}} = 1-\gamma_p s,\ c_r^{\mathrm{EO}} = 1 - \frac{1-\gamma_r s}{1-\gamma_p s}\gamma_p s, \quad \pi_{\mathrm{EO}}(s) = \frac{\alpha_p\gamma_p s}{1-\gamma_p s} + \frac{\alpha_r\gamma_p s}{1-\gamma_p s} - \tau(s);$$

(3) When $\beta_p > \alpha_p + \frac{1-\gamma_p s}{1-\gamma_r s}\alpha_r$,

$$c_p^{\mathrm{ET}} = c_r^{\mathrm{ET}} = 1 - \gamma_p s, \qquad \pi_{\mathrm{ET}}(s) = \frac{\alpha_p\gamma_p s}{1-\gamma_p s} + \frac{\alpha_r\gamma_p s}{1-\gamma_r s} - \tau(s);$$

$$c_p^{\mathrm{EO}} = 1-\gamma_p s, c_r^{\mathrm{EO}} = 1 - \frac{1-\gamma_r s}{1-\gamma_p s}\gamma_p s, \qquad \pi_{\mathrm{EO}}(s) = \frac{\alpha_p\gamma_p s}{1-\gamma_p s} + \frac{\alpha_r\gamma_p s}{1-\gamma_p s} - \tau(s).$$

In the main paper, we focus only on the most interesting case (1) when the thresholds are in the overlapping regions of all the distributions under both equal treatment and equal opportunity, and assume $\beta_p \leq \alpha_r + \alpha_p$. Our results continue to hold in other cases too (see Appendix).

4. Comparisons

In the previous section, for each of the two fairness constraints, we have obtained the optimal thresholds for a given level of learning s and converted the profit into a function of only s. This allows us to explore the relationship between the profit and the learning under different constraints, and examine how the corresponding optimal decisions influence the decision-maker and the applicants. We present the results in this section.

4.1. Learning

Let s^{ET} and s^{EO} be the optimal learning under equal treatment and equal opportunity, respectively. We compare the two quantities to find which policy leads to more learning. Before we formally present and prove our result, it is convenient to show the following lemma first.

We now present Theorem 1 which compares the optimal learning levels under the two constraints.

Theorem 1 *The optimal learning effort is higher under equal treatment than under equal opportunity, i.e.,* $s^{ET} \geq s^{EO}$.

This theorem suggests that equal opportunity discourages learning effort compared to equal treatment. To understand the intuition, we separate each profit function into two parts: revenue, denoted as $R(s)$, and cost, $\tau(s)$.

$$\pi_{\text{ET}}(s) = R_{\text{ET}}(s) - \tau(s), \text{ where } R_{\text{ET}}(s) = \frac{\alpha_p \gamma_r s}{1-\gamma_p s} + \frac{\alpha_r \gamma_r s}{1-\gamma_r s} + \frac{\beta_p \gamma_p s}{1-\gamma_p s} - \frac{\beta_p \gamma_r s}{1-\gamma_p s};$$
$$\pi_{\text{EO}}(s) = R_{\text{EO}}(s) - \tau(s), \text{ where } R_{\text{EO}}(s) = \frac{\alpha_p \gamma_r s}{1-\gamma_r s} + \frac{\alpha_r \gamma_r s}{1-\gamma_r s} + \frac{\beta_p \gamma_p s}{1-\gamma_p s} - \frac{\beta_p \gamma_r s}{1-\gamma_r s}.$$

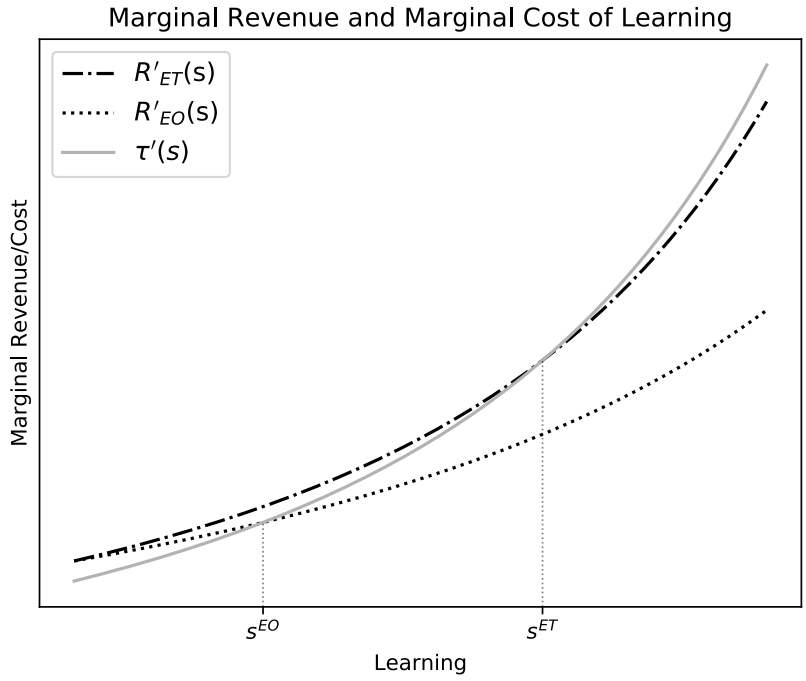

Figure 4 **Marginal cost of s (learning) and marginal revenue of s under different constraints: equal treatment(ET) and equal opportunity(EO). For any given s, $R'_{\text{ET}} \geq R'_{\text{EO}}$. Therefore, the optimal values of learning, at which marginal costs equal marginal revenues, have the order $s^{\text{ET}} \geq s^{\text{EO}}$**

Intuitively, when we switch from equal treatment to equal opportunity, the marginal benefit of learning is reduced. Equal opportunity requires a lower acceptance threshold for the protected group than for the regular group (see Proposition 1), while equal treatment requires the same thresholds. When the decision-maker makes additional learning effort, it better separates bad candidates from good candidates, and can therefore lower the thresholds to accept more candidates. Under equal treatment, the decision-maker always lower the two thresholds by the same amount. Under equal opportunity, the decision-maker has to lower the threshold for the protected group more than the

threshold for the regular group, because equal opportunity requires the same rate of acceptance for good candidates, while the distribution of the protected good candidates is more spread out due to the less efficient learning. As a result, the decision-maker either suffers from a higher loss brought by the additional protected bad candidates whom it has to accept, or fails to achieve a higher profit provided by some regular good candidates whom it could accept otherwise. In other words, equal opportunity restricts the decision-maker to partially realizing the benefit of learning, and eventually leads to a lower optimal learning effort.

The "cost of fairness" has been well recognized in the fair machine learning literature – when we enforce fairness constraints, the prediction accuracy usually decreases. Current view of the trade-off between fairness and accuracy, however, is instantaneous. Previous literature mostly focuses on a machine learning model with fixed predictive power, and compares the prediction accuracy for the model under certain fairness constraints. This accuracy loss corresponds to the case when s is fixed and the decision-maker only chooses acceptance thresholds in accordance to different constraints in our model. However, the "cost of the fairness" is more than this direct impact on accuracy. We highlight that equal opportunity effectively act as a "tax" on learning and reduce the marginal benefit of learning. When the benefit of learning is discounted, the decision maker is less incentivized to put effort into better prediction, which further decreases prediction accuracy.

4.2. Impact on the Decision Maker

As a business entity, the decision-maker's primary goal is to maximize its profit. Let π^*_{ET} and π^*_{EO} be the optimal profits under equal treatment and equal opportunity, respectively. We compare them in the following proposition.

Proposition 2 *The profit of the decision maker is higher under equal treatment than under equal opportunity, i.e.,* $\pi^*_{ET} \geq \pi^*_{EO}$.

As we haves shown in the previous section, under equal treatment the decision-maker is able to take more advantage of learning, while the cost of learning remains the same regardless of the fairness constraint. Therefore, the decision-maker's profit is higher under equal treatment. More specifically, with the same amount of learning s, the decision-maker has to set a lower threshold for the protected users under equal opportunity, which results in more acceptance of both protected-bad candidates and protected-good candidates. Because $\beta_p \geq \alpha_p$, the loss from the additional bad candidates is higher than the profit form the additional good candidates, and it leads to net loss in profit under equal opportunity. Furthermore, the decision-maker optimally chooses a higher amount of learning s under equal treatment, which means more learning can bring more profits. Therefore, overall the decision-maker's profit is higher under ET compared to EO.

4.3. Impact on the Regular Group

Candidates care about getting accepted. All candidates want to be accepted, but not all candidates are the same: some will turn out to be good, others bad. We call the acceptance of a good candidate a *successful acceptance*, and the acceptance of a bad candidate a *failed acceptance*. It is clear that a successful acceptance is beneficial for candidates too. For example, in the case of loan granting, the fact that these candidates applied for the loan and eventually paid it back suggests that they are able to use the loan in meaningful ways that increase their utilities. From a social planner's perspective, successful loans are well justified because their recipients are people who deserve the loan. What is less obvious is that failed acceptance can be harmful for candidates. When a bad candidate is accepted, not only does the decision-maker suffer from a financial loss, the candidate also loses trust – his or her score will decrease and it will be more difficult to get opportunities that require a good score in the future (Liu et al. 2018). From a group perspective, failed acceptance has an opportunity cost, as it could be better utilized if given to the good candidates. Moreover, if more bad candidates are accepted, the observed proportion of bad candidates of the group would increase, which may leave the impression that this group is "riskier" when it is actually just the poor selection of candidates. While it may be hard to conclude whether failed acceptance is beneficial or harmful to the candidates in an absolute sense, it is obvious that failed acceptance is less beneficial than successes. When examining the effect of the fairness notions on the regular group (or the protected group), we focus on two statistics:

1. The fraction of good candidates who get accepted, which equals to the number of successful acceptance divided by the total number of good candidates in the group. We call it *coverage rate*, as in the rate of the deserving candidates (good candidates) in the group being served (covered). This is technically equivalent to the True Positive Rate that is widely used in machine learning literature. We denote it as ϕ_A^C, where $A \in \{p,\ r\}$ is the group membership (Protected or Regular) and $C \in \{ET,\ EO\}$ is the fairness constraint (Equal Treatment or Equal Opportunity).
2. The fraction of successful acceptance in all the accepted candidates for the group, which equals to the number of successful acceptance divided by the total number of accepted candidates to the group. We call it *success rate*, as it measures the rate of success among all the accepted candidates. It is denoted as δ_A^C, where A is the group membership and C is the fairness constraint.

Based on the previous reasoning, these two rates are strong indicators of the group welfare and are positively correlated with it. They allow us to compare group welfare under different fairness constraints without assuming a specific relationship between the utility of a successful acceptance and the utility of a failed acceptance for the group. We first examine the coverage rate of the regular group:

Proposition 3 *For the regular group, a higher fraction of good candidates would be accepted under equal treatment than under equal opportunity, i.e.,* $\phi_r^{ET} \geq \phi_r^{EO}$.

Under equal opportunity, the optimal amount of learning is lower (Theorem 1), so the decision-maker is less certain about quality of candidates. Consequentially, it sets a higher acceptance threshold. Moreover, with a lower learning, there are fewer regular-good candidates with scores higher than any given threshold. Therefore, a smaller fraction of good candidates in the regular group gets accepted under equal opportunity compared to under equal treatment. As the total number of good candidates in the regular group remains the same, this also means that fewer good candidates in the regular group are accepted under equal opportunity.

With the optimal thresholds, the decision-maker never accepts bad candidates in the regular group. In other words, the success rate of the regular group is 1 both under equal treatment and under equal opportunity. Combined with Proposition 3, it means that overall fewer candidates from the regular group are accepted under equal opportunity. Therefore, compared to equal treatment, equal opportunity makes the regular group worse off.

4.4. Impact on the Protected Group

Now we move to the impact of the fairness constraints on the protected group. As mentioned in the previous section, we compare the protected group's welfare under equal treatment and under equal opportunity by examining two statistics of the group: coverage rate (the fraction of good candidates who get accepted) and success rate (the fraction of successful acceptance in all the accepted candidates).

Equal opportunity is designed to protect the protected group in terms of the coverage rate. In most of the cases, the protected group would have a lower coverage rate if we apply the same thresholds to both groups. For example, the case study in Hardt et al. (2016) shows that if we grant loans based on candidates' FICO scores, then under race-blind thresholds, Hispanic and black people who would not default are granted loans at much lower rates than others. Similarly, Angwin et al. (2016) point out that black people who did not re-offend were less likely to be labeled as low risk by COMPAS[6] than white people who did not re-offend. Therefore, if we release defendants with COMPAS risk scores below a certain threshold, the black "good defendants" (those who would not re-offend) will have a smaller chance to be released compared to the white "good defendants." Our model provides one explanation for this phenomenon: Because the learning efficiency is lower for the protected group, the decision-maker always learns more about the regular group. This means bad candidates and good

[6] A risk score produced by an algorithm that predicts recidivism.

candidates are always better separated in the regular group. As good candidates' scores are more concentrated in high values with better separation, more regular good candidates than protected good candidates will have scores above any given threshold. Therefore, under equal treatment, the regular group always has a higher coverage rate.

As equal opportunity requires the same coverage rate for the two groups, intuitively it should result in a higher coverage rate for the protected group compared to the coverage rate under equal treatment. Indeed, with the same amount of learning, the decision-maker has to choose a lower threshold for the protected group under equal opportunity than under equal treatment. This lower threshold forces the decision-maker to accept more protected candidates (both bad candidates and good candidates), and thus leads to a higher coverage rate under equal opportunity. We call it *threshold effect* of equal opportunity. However, there is a second force that also influences the coverage rate. Under equal opportunity, the optimal amount of learning is lower. The decision-maker is less certain about the quality of candidates, and would raise the thresholds. Meanwhile, with lower learning, fewer protected good candidates would have scores above any given threshold. Just as less learning under equal opportunity leads to a lower coverage rate for the regular group, it also decreases the coverage rate for the protected group. We call it *learning effect* of equal opportunity. As threshold effect and learning effect are two opposing forces, the protected group's coverage rate under equal opportunity could be either higher or lower, depending on the size of the two forces. Specifically, when learning effect is stronger than threshold effect, fewer protected good candidates would be accepted under equal opportunity. In other words, equal opportunity can harm the very group it aims to protect.

Clearly, the cost function $\tau(s)$ affects the size of threshold effect and learning effect. To analytically show that the protected group can have a lower coverage rate under equal opportunity, we assume a cost function, $\tau(s) = \frac{ks}{1-\gamma_r s}$, where k is a parameter that characterizes the level of the learning cost. When k is too high, i.e., it is too costly to learn, the optimal learning would be 0 and the decision-maker would not accept any candidates. In this case, the market fails under both equal treatment and equal opportunity. When k is too low, i.e., it is too easy to learn, the decision maker would choose to perfectly separate the good candidates from the bad candidates at least for the regular group, which we assume could never happen. Thus, we assume that $k \in [\alpha_r\gamma_r + \rho^2(\alpha_p\gamma_r - \beta_p\gamma_r + \beta_p\gamma_p),\ \alpha_r\gamma_r + \alpha_p\gamma_r - \beta_p\gamma_r + \beta_p\gamma_p)$, where $\rho = \frac{1}{2-\frac{\gamma_p}{\gamma_r}}$. The value of k in this range ensures that neither market failure nor perfect separation happens, i.e., $0 < s^{\mathrm{ET}}, s^{\mathrm{EO}} < \frac{1}{2\gamma_r}$.

We assume k is within this range and focus on cases where separations of the bad candidates and the good candidates are positive but not perfect. Note that what matters here is the size of k

relative to other parameters, instead of the absolute value of k. Therefore, we can re-write k in other parameters and a scale parameter σ:

$$k = \alpha_r\gamma_r + \sigma^2(\alpha_p\gamma_r - \beta_p\gamma_r + \beta_p\gamma_p), \quad \sigma \in [\rho, 1).$$

The scale parameter σ represents the actual level of learning cost: when $\sigma < \rho$, the decision-maker would choose to perfectly separate the bad candidates and the good candidates because the learning cost is minimal; as σ increases, the optimal amounts of learning decrease as learning becomes more costly; when $\sigma \geq 1$, the market fails because it is too expensive to learn.

Theorem 2 *For the protected group, a higher fraction of good candidates would get accepted under equal treatment than under equal opportunity when the ratio of expected profit (α_p) to expected loss (β_p) is sufficiently low for the protected group. That is, $\phi_p^{ET} \geq \phi_p^{EO}$, if $\frac{\alpha_p}{\beta_p} \leq 1 - \frac{1-(2-\sigma)^2)\sigma^2}{(1-\sigma^2)(2-\sigma)^2} \cdot \frac{\gamma_p}{\gamma_r}$.*

As previously discussed, equal opportunity would harm the protected group in its coverage rate when the learning effect is greater than the threshold effect. While the two effects are interdependent and determined by all the parameters jointly, the threshold effect is mostly affected by the learning efficiency gap $(\gamma_r - \gamma_p)$, and the learning effect is influenced more by the expected loss from the protected group $(\beta_p - \alpha_p)$. Imagine we start from the optimal learning and thresholds under equal treatment, and change the fairness constraint to equal opportunity. Now the decision-maker has to set a lower threshold for the protected group to match its coverage rate to the regular group's coverage rate. The higher the learning efficiency gap is, the more the decision-maker has to lower the threshold. Since the expected profit from the protected group is negative $(\beta_p > \alpha_p)$, the decision-maker suffers from a loss when it lowers the threshold and accepts more protected candidates. To mitigate the loss, the decision-maker would decrease the learning effort, as this reduces both the total cost of learning and the difference of the separations in the protected group and in the regular group. The higher the expected loss from the protected group $(\beta_p - \alpha_p)$, the more the decision-maker suffers from a lower threshold, and thus the more it is motivated to reduce the learning effort. In the extreme case, if the expected loss from the protected group is 0, the decision-maker would be indifferent to setting a lower threshold, and has no incentive to reduce learning, which means no learning effect. Therefore, when the expected loss from the protected group is sufficiently large relative to the learning efficiency gap, the learning effect dominates the thresholds effect, and equal opportunity would result in a lower coverage rate for the protected group compared to equal treatment.

After showing that the coverage rate of the protected group can be lower under equal opportunity than under equal treatment, we now compare the success rate of the protected group under the two fairness constraints.

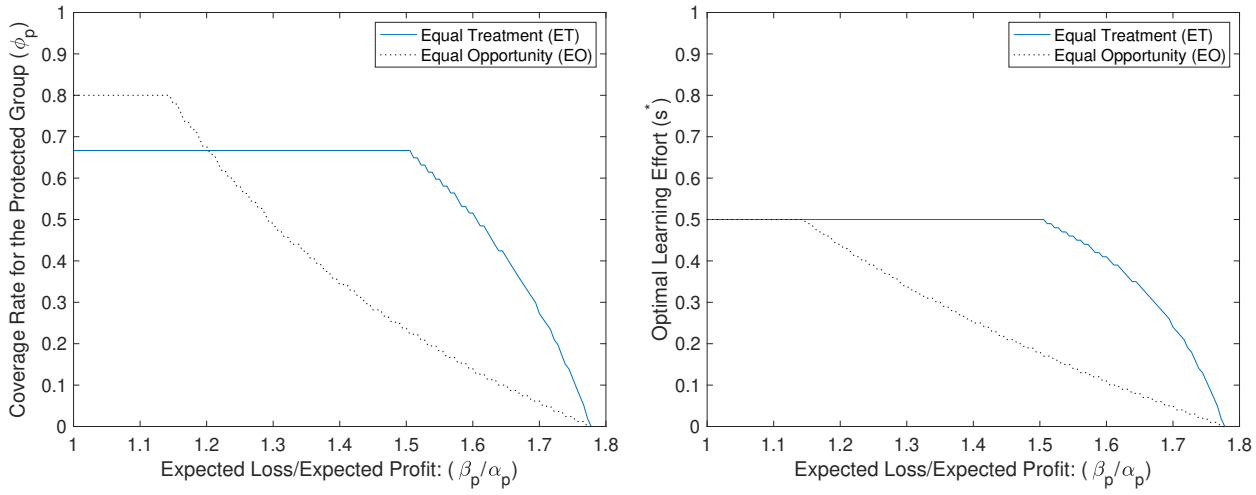

Figure 5 **The figure on the left depicts that in a market with a high ratio of expected loss to expected profit, the fairness notion of equal opportunity can hurt the protected group. The figure on the right explains the intuition behind this result. In a risky market in which the ratio of expected loss to expected profit is high, firms have low incentive to exert effort in learning.**

Theorem 3 *For the protected group, a higher fraction of the approved candidates are good candidates under equal treatment than under equal opportunity, i.e.,* $\delta_p^{ET} > \delta_p^{EO}$.

With any degree of learning and separation, the bad candidates' scores are more concentrated in low values while the good candidates' scores are more concentrated in high values. Therefore, as the decision-maker lowers the threshold for the protected group, the additionally approved candidates will have a lower (or at best, the same) percentage of good candidates, which decreases the overall success rate. Moreover, under equal opportunity, the learning effort is lower, thus the bad candidates and the good candidates are less separated in general, which also contributes to a lower success rate.

Theorem 2 suggests that fewer good candidates in the protected group would get accepted under equal opportunity when the expected loss from the group is high, while theorem 3 says among the approved protected candidates, the portion of good candidates is always smaller under equal opportunity. This means very often equal opportunity would make the protected group worse off as it leads to lower values in both coverage rate and success rate. Even when the decision-maker approves more protected good candidates under equal opportunity because the expected loss from the protected group is low, the additional good candidates that are approved always come at the cost of even more additional bad candidates who also have to be approved. Overall, equal opportunity can harm the protected group–the very group it aims to protect.

5. A General Fairness Notion

Most fairness notions, including equal opportunity and other definitions such as demographic parity, equalized odds and conditional statistical parity,[7] aim to achieve equal impact and, hence, require the decision maker to *lower* the threshold for the protected group. Corbett-Davies et al. (2017) show that for several popular definitions of fairness, the utility maximizing algorithms require group specific thresholds, and when the protected group is at a disadvantage, this means a lower threshold for it. Indeed, many algorithms that aim to achieve different fairness notions involve setting a lower threshold for the protected group either directly (e.g., Hardt et al. (2016)) or indirectly through data processing and transformation (e.g., Calders et al. (2009)). In this section, we show that our main results continue to hold of a general fairness notion having this property. First we define such a general fairness notion. Define

$$\theta = \frac{1-c_p}{1-c_r}. \tag{12}$$

A fairness notion can be simply specified by the value of θ. A value of $\theta > 1$ means that the fairness notion requires the decision maker to lower the threshold for the protected group. When $\theta = 1$, this fairness notion is equivalent to equal-treatment. Similarly, when $\theta = \frac{1-\gamma_p s}{1-\gamma_r s}$, this fairness notion is equivalent to equal opportunity.

Similar to Section 4, we focus on the case when $c_r = 1-\gamma_r s$. Using the definition of θ, we have $c_p = 1-\theta(1-c_r)$. Thus, we have $1-c_r = \gamma_r s$ and $1-c_p = \theta(1-c_r) = \theta\gamma_r s$. Let π_θ represent the profit of the decision maker under this general fairness notion. Then, π_θ can be written as

$$\pi_\theta = \frac{\gamma_r s \alpha_r}{1-\gamma_r s} + \frac{\theta\gamma_r s \alpha_p}{1-\gamma_p s} - \frac{(\theta\gamma_r s - \gamma_p s)\beta_p}{1-\gamma_p s} - \tau(s).$$

We now analyze the effect of the general fairness notion defined in this section on the learning effort (s) exerted by the decision-maker. Define $G_1(\theta) = \gamma_p\beta_p - \theta\gamma_r(\beta_p - \alpha_p)$. The above expression of profit can now be written as

$$\pi_\theta = \frac{\gamma_r s \alpha_r}{1-\gamma_r s} + \frac{sG_1(\theta)}{1-\gamma_p s} - \tau(s). \tag{13}$$

Taking the derivative of π_θ with respect to s, we get

$$\frac{d\pi_\theta}{ds} = \frac{\gamma_r \alpha_r}{(1-\gamma_r s)^2} + \frac{G_1(\theta)}{(1-\gamma_p s)^2} - \tau'(s). \tag{14}$$

Using the above expression, it is easy to see that $\frac{d}{d\theta}\left(\frac{d\pi_\theta}{ds}\right) \leq 0$, because $\frac{dG_1(\theta)}{d\theta} \leq 0$. Let $s^*(\theta)$ be the optimal learning effort by the decision-maker. Then, we have the following result

[7] Demographic parity requires equal acceptance rate for both groups; equalized odds requires equal true positive rate and equal false positive rate; conditional statistical parity requires equal acceptance rate conditioning on a set of risk factors.

Theorem 4 *The optimal learning effort $s^*(\theta)$ decreases with θ.*

The proof and the intuition of the above theorem is similar to that of Theorem 1. We now proceed to analyze the impact of the general fairness notion on the profit of the decision-maker. At the optimal level of learning effort s^*, we can write the profit of the decision-maker as

$$\pi_\theta(s^*) = \frac{\gamma_r s^* \alpha_r}{1-\gamma_r s^*} + \frac{s^* G_1(\theta)}{1-\gamma_p s^*} - \tau(s^*). \tag{15}$$

Taking the derivative with respect to θ, we get

$$\frac{d\pi_\theta(s^*)}{d\theta} = \left[\frac{\gamma_r \alpha_r}{(1-\gamma_r s^*)^2} + \frac{G_1(\theta)}{(1-\gamma_p s^*)^2} - \tau'(s^*)\right]\frac{ds^*}{d\theta} + \frac{s^*}{1-\gamma_p s^*}\frac{dG_1(\theta)}{d\theta}. \tag{16}$$

Using (14), we have

$$\frac{d\pi_\theta(s^*)}{d\theta} = \frac{d\pi_\theta}{ds}\Big|_{s=s^*} \times \frac{ds^*}{d\theta} + \frac{s^*}{1-\gamma_p s^*}\frac{dG_1(\theta)}{d\theta}. \tag{17}$$

As s^* is the optimal learning effort, it should satisfy

$$\frac{d\pi_\theta}{ds}\Big|_{s=s^*} = 0,$$

if it is an interior solution. Thus, we have

$$\frac{d\pi_\theta(s^*)}{d\theta} = \frac{s^*}{1-\gamma_p s^*}\frac{dG_1(\theta)}{d\theta}. \tag{18}$$

It is easy to see that $\frac{d\pi_\theta(s^*)}{d\theta} \leq 0$, because $\frac{dG_1(\theta)}{d\theta} \leq 0$. Formally, we have the following result.

Proposition 4 *The profit of the decision-maker decreases with θ.*

Figure 6, shows that the profit of the decision-maker decreases with an increase in θ. Also, for a given value of θ, the profit decreases as the market becomes more risky (as β_p/α_p increases). When the market is too risky, the decision-maker's profit goes down to zero and it ceases to exist. Under a lower value of θ, the decision maker can continue to operate for a much riskier market, with a positive profit.

We now proceed to analyze the effect of the general fairness notion on the welfare of the regular group. Specifically, we analyze the impact of the general fairness notion on the coverage rate for the regular group. We can write the coverage rate for regular group as

$$\phi_r(\theta) = \frac{1-c_r}{1-\gamma_r s^*(\theta)} = \frac{\gamma_r s^*(\theta)}{1-\gamma_r s^*(\theta)}.$$

From Theorem 4 we know that $s^*(\theta)$ decreases with θ. Thus, it is easy to see that the coverage rate for the regular group, i.e., $\phi_r(\theta)$, decreases with θ. Formally, we have the following result

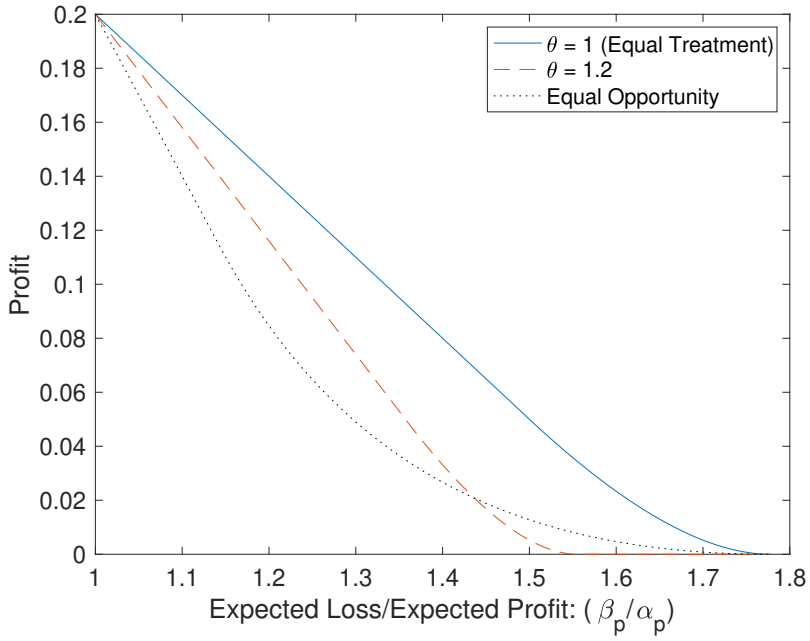

Figure 6 **The profit of the decision-maker decreases with an increase in θ. Also, for a given value of θ, the profit decreases as the market becomes more risky (as β_p/α_p increases). When the market is too risky, the decision-maker's profit goes down to zero and it ceases to exist. Under a lower value of θ, the decision-maker can continue to operate for a much riskier market, with positive profit.**

Proposition 5 *The coverage rate for the regular group, i.e., $\phi_r(\theta)$, decreases with θ.*

We now proceed to analyze the impact of the general fairness notion on the coverage rate for the protected group. We know that

$$\phi_p(\theta) = \frac{1-c_p}{1-\gamma_p s^*(\theta)} = \frac{\theta\gamma_r s^*(\theta)}{1-\gamma_p s^*(\theta)} \ . \tag{19}$$

The general fairness notion will hurt the protected group if it leads to a lower coverage rate for the protected group compared to the coverage rate under equal treatment fairness notion, i.e., $\phi_p(\theta) \leq \phi_p(1)$. This condition simplifies to

$$\frac{\beta_p}{\alpha_p} \geq \hat{r}, \tag{20}$$

where $\hat{r} = \frac{\gamma_r[A_2(\theta)-\theta B_2(\theta)]}{(\gamma_r-\gamma_p)A_2(\theta)-(\theta\gamma_r-\gamma_p)B_2}$, $A_2(\theta) = \frac{\theta^2(s^*(\theta))^2}{(s^*(1))^2}$, and $B_2(\theta) = \frac{1-\gamma_r s^*(\theta)}{1-\gamma_r s^*(1)}$. Formally, we have the following result

Theorem 5 *The general fairness notion will hurt the protected group if the market is riskier than $\hat{r}$. That is, $\phi_p(\theta) \leq \phi_p(1)\ \forall\ \theta \geq 1$, if $\frac{\beta_p}{\alpha_p} \geq \hat{r}$.*

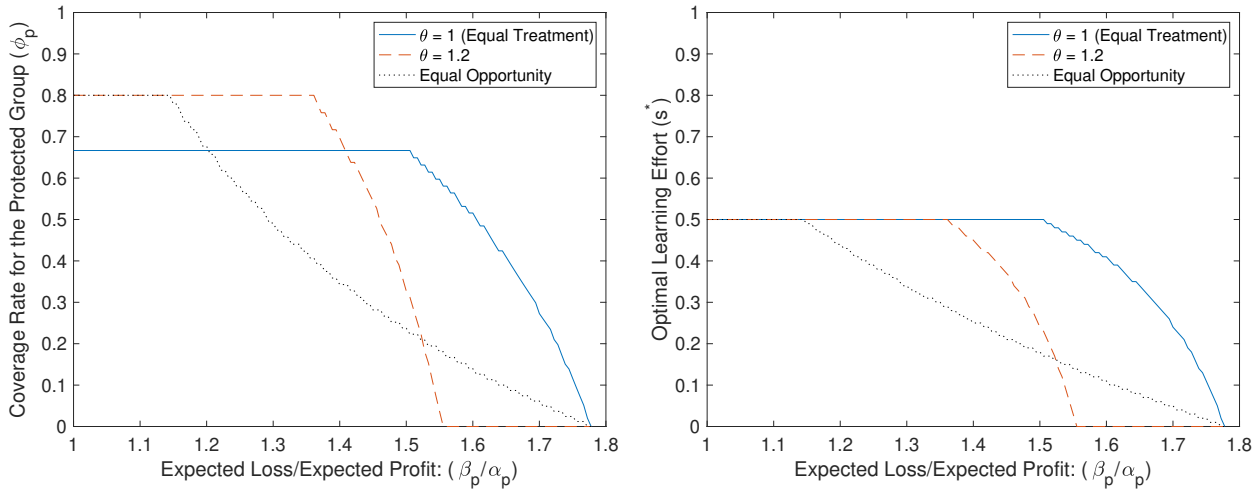

Figure 7 **The figure on the left illustrates that when the market is risky (i.e., when β_p/α_p is high), the coverage rate for the protected group is highest when $\theta = 1$, i.e., under equal treatment. In other words, a general fairness notion, which requires $\theta > 1$, can eventually hurt the protected group in a risky market. The figure on the right gives the intuition of this result. Under a risky market, firms have lower incentive to learn.**

The left-hand-side of Figure 7 plots the coverage rate ϕ_p for $\theta = 1,\ 1.2$ and equal opportunity. We see that when the market is risky, i.e., when β is high, equal opportunity and all higher values of θ lead to a lower coverage rate compared to $\theta = 1$ (equal treatment). In other word, in a risky market a fairness notion that aims to lower the threshold for the protected group can actually hurt the protected group. The right-hand side of Figure 7 gives the intuition behind the result. This figure plots the learning effort $s^*(\theta)$ for $\theta = 1,\ 1.2$ and equal opportunity.

6. Robustness Check Under Additional Features

In this section, we check the robustness of our base model by adding the following features:

- **Two-Period Model:** Our base model is a single period model. We extend this to a dynamic two-period model and show that our results continue to hold for a dynamic setting.
- **Same Default Rate for Both Groups:** In our base model, we assumed that the default rate for the protected group is higher than that of the regular group (i.e., $d_p > d_r$). We now relax this assumption and show the robustness of our results when the default rate is the same for both the groups ($d_p = d_r = d$).
- **Different Population Sizes:** We also relax the assumption that both protected and regular groups are of equal size and show the validity of our results when one group is larger than the other.

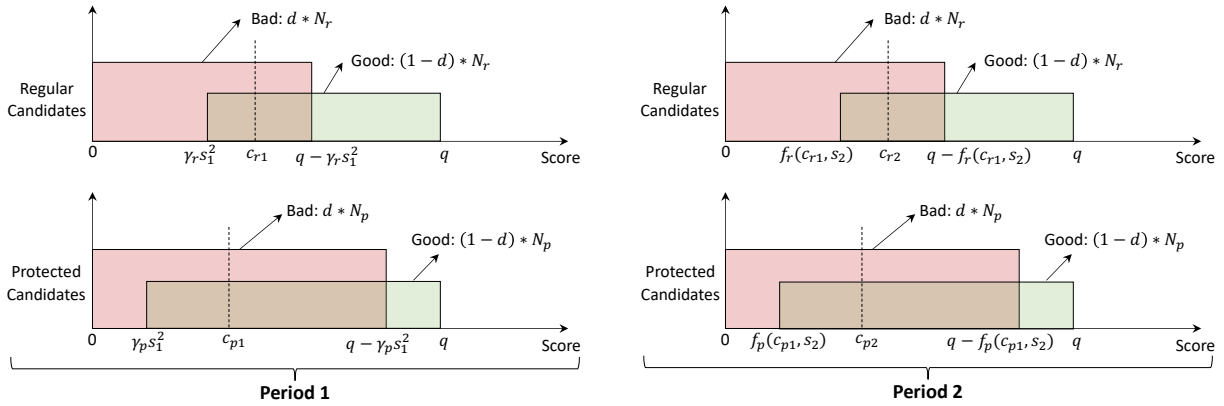

Figure 8 **A dynamic two-period model: The learning rate in the second-period depends on the cut-offs chosen in the first period. Specifically, a lower cut-off in the first period (more accepted applicants) leads to a higher learning rate in the next period.**

- **Non-Linear Learning:** Our base model assumes that the separation of "good" and "bad" applicants happens at a linear rate (γ_p and γ_r for the protected and regular groups, respectively). We now show that our results continue to hold for a non-linear learning rate.
- **Competition:** Our base model is a monopoly setting. We extend this to a competitive setting and show the robustness of our main results.

Figure 8 pictorially depicts the two-period model. The decision-maker dynamically makes decisions in two periods. The left sub-figure in Figure 8 represents the distribution of good and bad applicants in both the groups in the first period. Similarly, the sub-figure on the right represents this distribution in the second period. Specifically, s_1 and s_2 represents the learning effort exerted by the decision-maker in the first and second periods respectively. Similarly, the cut-offs chosen in the first and second period are represented by $\{c_{r1},\ c_{r2}\}$ for regular applicants and by $\{c_{p1},\ c_{p2}\}$ for protected applicants.

In Figure 8, we also note that the rate of separation of good and bad applicants in the first period depend on the learning effort (s_1) in a quadratic fashion ($\gamma_r s_1^2$ and $\gamma_p s_1^2$). More importantly, in the second period, the rate of separation of good and bad applicants is represented by $f_r(c_{r1}, s_2)$ and $f_p(c_{p1}, s_2)$ for regular and protected applicants, respectively. Note that these rates $\big(f_r(c_{r1}, s_2)$ and $f_p(c_{p1}, s_2)\big)$ depend on the cut-offs chosen in the previous period (c_{r1} and c_{p1}). Specifically, we use the following functional forms for tractability:

$$f_r(c_{r1}, s_2) = q - c_{r1} + \gamma_r s_2^2 \tag{21}$$

and

$$f_p(c_{p1}, s_2) = q - c_{p1} + \gamma_p s_2^2. \tag{22}$$

Note that a lower cut-off in the first period help in better separation in the second period. This captures the idea that more applicants accepted in the current period helps in better identification of good applicants in the future (similar to the exploration-exploitation trade-off).

Also note that the fraction of bad applicants is the same in both groups (d). The number of regular applicants is N_r and there are N_p protected applicants (in the base model N_p and N_r were normalized to 1). The scores of applicants is now distributed between 0 and q. We analyze the general fairness notion discussed in Section 5. That is, the cut-offs chosen by the decision-maker should satisfy the following constraints corresponding to two periods:

$$\frac{q - c_{p1}}{q - c_{r1}} = \theta, \tag{23}$$

and

$$\frac{q - c_{p2}}{q - c_{r2}} = \theta. \tag{24}$$

Let $\vec{c} = (c_{r1}, c_{r2}, c_{p1}, c_{p1},)$ and $\vec{s} = (s_1, s_2)$. The profit of the decision-maker can be written as follows:

$$\pi(\vec{c}, \vec{s}) = \pi_1 + \pi_2, \tag{25}$$

where π_1 is the profit in the first period and π_2 is the profit in the second period. Let $\bar{d} = 1 - d$. Using Figure 8 we have

$$\begin{aligned}\pi_1 = &\frac{(q - c_{r1})\bar{d}\alpha N_r}{q - \gamma_r s_1^2} - \frac{(q - \gamma_r s_1^2 - c_{r1})d\beta N_r}{q - \gamma_r s_1^2} \\ &+ \frac{(q - c_{p1})\bar{d}\alpha N_p}{q - \gamma_p s_1^2} - \frac{(q - \gamma_p s_1^2 - c_{p1})d\beta N_p}{q - \gamma_p s_1^2} - \tau(s_1)\end{aligned} \tag{26}$$

and

$$\begin{aligned}\pi_2 = &\frac{(q - c_{r2})\bar{d}\alpha N_r}{q - f_r(c_{r1}, s_2)} - \frac{(q - f_r(c_{r1}, s_2) - c_{r2})d\beta N_r}{q - f_r(c_{r1}, s_2)} \\ &+ \frac{(q - c_{p2})\bar{d}\alpha N_p}{q - f_p(c_{p1}, s_2)} - \frac{(q - f_p(c_{p1}, s_2) - c_{p2})d\beta N_p}{q - f_p(c_{p1}, s_2)} - \tau(s_2).\end{aligned} \tag{27}$$

The decision-maker solves the following problem:

$$\begin{aligned}&\max_{\vec{c}, \vec{s}} \ \pi(\vec{c}, \vec{s}) \\ &\text{s.t. } \frac{q - c_{p1}}{q - c_{r1}} = \theta, \ \frac{q - c_{p2}}{q - c_{r2}} = \theta.\end{aligned}$$

We now proceed to the analysis of this general model. Similar to Section 4, we focus on the case when β is large enough such that the optimal cut-offs in both the periods are as follows:

$$c^*_{r1} = q - \gamma_r s_1^2,$$
$$c^*_{r2} = q - f_r(c^*_{r1}, s_2).$$

Using (23) and (61) we have

$$c^*_{p1} = q - \theta(q - c^*_{r1}),$$
$$c^*_{p2} = q - \theta(q - c^*_{r2}).$$

Thus, we have

$$q - c^*_{r1} = \gamma_r s_1^2,\ q - c^*_{r2} = \gamma_r(s_1^2 + s_2^2),$$
$$q - c^*_{p1} = \theta\gamma_r s_1^2,\ q - c^*_{p2} = \theta\gamma_r(s_1^2 + s_2^2).$$

Let $G_1(\theta) = \gamma_p d\beta - \theta\gamma_r(d\beta - \bar{d}\alpha)$. Substituting above values in π_1 and π_2, we get

$$\pi_1 = \frac{\gamma_r s_1^2 \bar{d}\alpha N_r}{q - \gamma_r s_1^2} + \frac{s_1^2 G_1(\theta) N_p}{q - \gamma_p s_1^2} - \tau(s_1), \tag{28}$$

$$\pi_2 = \frac{\gamma_r (s_1^2 + s_2^2) \bar{d}\alpha N_r}{q - \gamma_r (s_1^2 + s_2^2)} + \frac{(s_1^2 + s_2^2) G_1(\theta) N_p}{q - \gamma_p (s_1^2 + s_2^2)} - \tau(s_2). \tag{29}$$

As $\pi = \pi_1 + \pi_2$, we have $\frac{d\pi}{ds_1} = \frac{d\pi_1}{ds_1} + \frac{d\pi_2}{ds_1}$. Taking derivative with respect to s_1, we have

$$\frac{d\pi_1}{ds_1} = \frac{2\gamma_r \bar{d}\alpha s_1 q N_r}{(q - \gamma_r s_1^2)^2} + \frac{2G_1(\theta) s_1 q N_p}{(q - \gamma_p s_1^2)^2} - \tau'(s_1),$$

$$\frac{d\pi_2}{ds_1} = \frac{2\gamma_r \bar{d}\alpha s_1 q N_r}{\left(q - \gamma_r (s_1^2 + s_2^2)\right)^2} + \frac{2G_1(\theta) s_1 q N_p}{\left(q - \gamma_p (s_1^2 + s_2^2)\right)^2} - \tau'(s_2).$$

Note that $\frac{dG_1(\theta)}{d\theta} < 0$. Using this, it is easy to verify that $\frac{d}{d\theta}\left(\frac{d\pi_1}{ds_1}\right) < 0$ and $\frac{d}{d\theta}\left(\frac{d\pi_2}{ds_1}\right) < 0$. Thus, we have

$$\frac{d}{d\theta}\left(\frac{d\pi}{ds_1}\right) = \frac{d}{d\theta}\left(\frac{d\pi_1}{ds_1}\right) + \frac{d}{d\theta}\left(\frac{d\pi_2}{ds_1}\right) < 0.$$

Similarly, we have

$$\frac{d\pi}{ds_2} = \frac{d\pi_2}{ds_2},$$
$$= \frac{2\gamma_r \bar{d}\alpha s_2 q N_r}{\left(q - \gamma_r (s_1^2 + s_2^2)\right)^2} + \frac{2G_1(\theta) s_2 q N_p}{\left(q - \gamma_p (s_1^2 + s_2^2)\right)^2} - \tau'(s_2).$$

Again using $\frac{dG_1(\theta)}{d\theta} < 0$, it is easy to verify that $\frac{d}{d\theta}\left(\frac{d\pi}{ds_2}\right) < 0$. Formally, we have the following result.

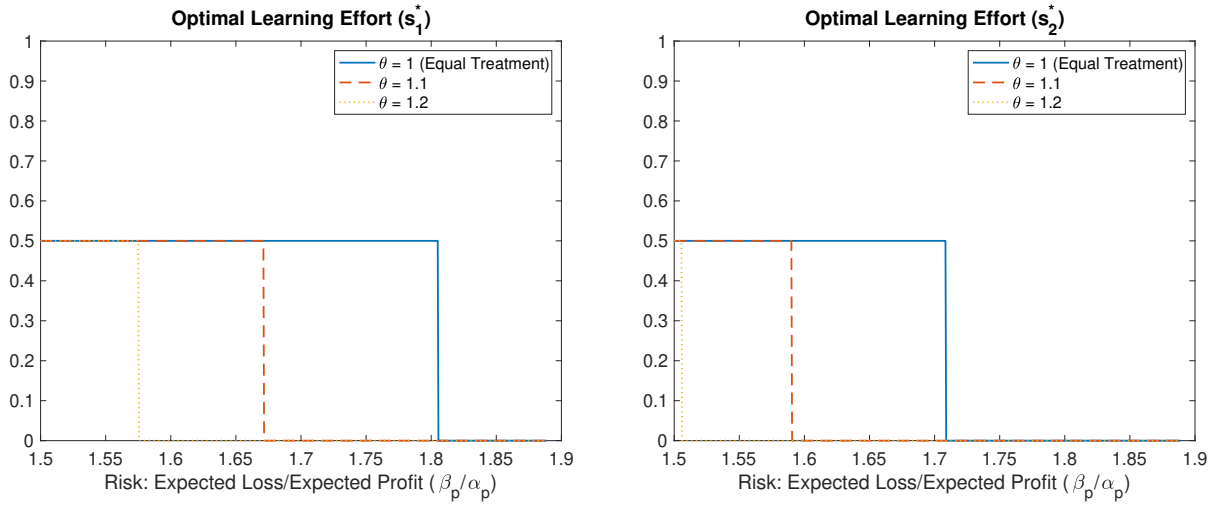

Figure 9 **Comparison of learning efforts: (i) As θ increases, the optimal learning efforts in both the periods, i.e., s_1^* and s_2^* decrease. (ii) The optimal learning effort in the first period is greater than or equal to that in the second period. Intuitively, this reflects the future-value of the learning effort in the first period.**

Theorem 6 *Optimal learning efforts in both periods decrease with θ. That is, $s_1^*(\theta)$ and s_2^* decrease with θ.*

The proof and the intuition of the above theorem is similar to that of Theorem 1. Figure 9 represents the optimal learning efforts exerted by the decision-maker in the first and the second period. The sub-figure on the left represents the optimal learning effort in the first period (s_2^*) and the sub-figure on the right represents the optimal learning effort in the second period (s_2^*). Note that (i) as θ increases, the optimal learning efforts in both the periods, i.e., s_1^* and s_2^* decrease, (ii) the optimal learning effort in the first period is greater than or equal to that in the second period. Intuitively, this reflects the future-value of learning effort in the first period.

Let $\phi_p(\theta)$ represent the fraction of accepted protected applicants. Then, we can write

$$\phi_p(\theta) = \frac{\theta\gamma_r s_1^{*2}}{q - \gamma_p s_1^{*2}} + \frac{\theta\gamma_r(s_1^{*2} + s_2^{*2})}{q - \gamma_p(s_1^{*2} + s_2^{*2})}.$$

Figure 10 plots $\phi_p(\theta)$ for different values of θ. We see that in a risky market, a higher value of θ can hurt the protected applicants.

Using 62, (63), and (64), we can obtain the optimal profit of the decision-maker. Figure 11 plots the profit of the decision-maker for several values of θ and risk in the market. We see that the profit of the decision-maker decreases with an increase in θ. The lower learning effort exerted by the decision-maker at a higher value of θ can be attributed to the lower profit that the decision-maker can expect to make as θ increases. In other words, the decision-maker doesn't earn enough to bear

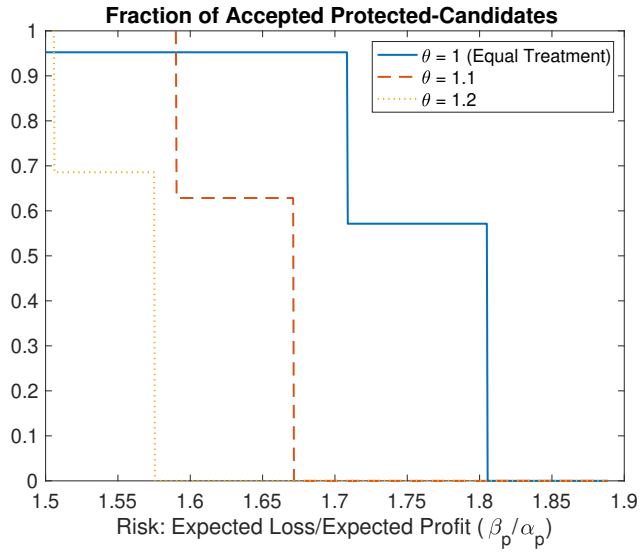

Figure 10 **In a risky market, fewer protected applicants are accepted as θ increases.**

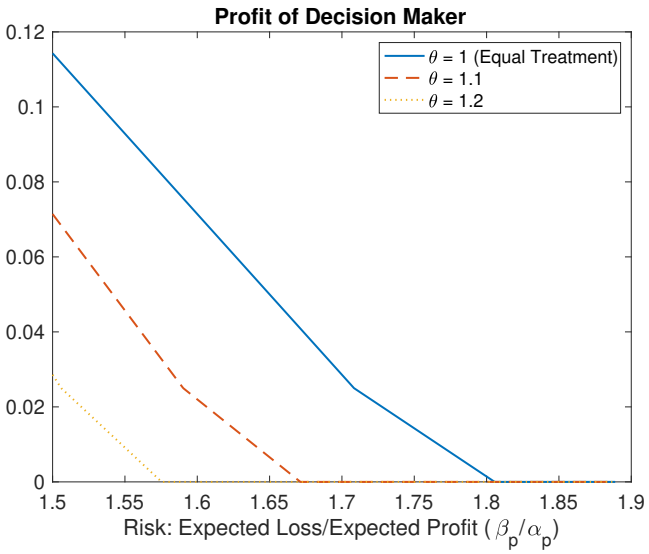

Figure 11 **Profit of the decision-maker decreases with θ.**

the cost of the learning effort. An increase in θ represents that the fairness constraint is becoming more stringent.

Learning Rate Dependent on Number of Selected Applicants: Recall that in (21) and (22), the second-period learning rates, i.e., $f_r(c_{r1}, s_2)$ and $f_p(c_{p1}, s_2)$, depend on the cut-offs chosen in the first period (c_{r1} and c_{p1}). A lower cut-off in the first period leads to a higher learning rate in the second-period, reflecting the future value of accepting more applicants in the beginning. Another approach to model the future value of accepting more applicants in the beginning is to make these learning rates depend directly on the number of accepted applicants in the first period. However,

this approach makes the model intractable. Thus, we now numerically analyze the two-period model of this section, when the second-period learning rates directly depend on the number of accepted applicants in the first period rather than on the cut-offs in the first period. Specifically, we now assume the following function form for the second-period learning rates:

$$f_r(n_{r1}, s_2) = \lambda n_{r1} + \gamma_r s_2^2 \tag{30}$$

and

$$f_p(n_{p1}, s_2) = \lambda n_{p1} + \gamma_p s_2^2, \tag{31}$$

where n_{r1} and n_{p1} are the number of regular and protected applicants selected in the first period, and λ is a parameter representing the magnitude of learning carried over from one period to another. Apart from these two new functional forms, the rest of the model in this section stays the same. Thus, we have

$$n_{r1} = \frac{(q - c_{r1})\bar{d}}{q - \gamma_r s_1^2} + \frac{(q - \gamma_r s_1^2 - c_{r1})d}{q - \gamma_r s_1^2} \tag{32}$$

and

$$n_{p1} = \frac{(q - c_{p1})\bar{d}}{q - \gamma_p s_1^2} + \frac{(q - \gamma_p s_1^2 - c_{p1})d}{q - \gamma_p s_1^2}. \tag{33}$$

Again assuming that β is large enough, such that we have

$$c_{r1}^* = q - \gamma_r s_1^2,$$

$$c_{p1}^* = q - \theta\gamma_r s_1^2,$$

$$c_{p2}^* = q - \theta f_r(n_{r1}, s_2),$$

$$c_{r2}^* = q - f_r(n_{r1}, s_2).$$

Let π_1 be the profit in the decision-maker in the first period. Then, we have

$$\pi_1 = \frac{\gamma_r s_1^2 \bar{d}\alpha}{q - \gamma_r s_1^2} + \frac{s_1^2 G_1(\theta)}{q - \gamma_p s_1^2} - \tau(s_1), \tag{34}$$

where $G_1(\theta) = f_p(n_{p1}, s_2)d\beta - \theta(d\beta - \bar{d}\alpha)f_r(n_{r1}, s_2)$. Similarly, let π_2 represent the profit in the second-period. Then, we have

$$\pi_2 = \frac{f_r(n_{r1}, s_2)\bar{d}\alpha}{q - f_r(n_{r1}, s_2)} + \frac{G_f(\theta)}{q - f_p(n_{p1}, s_2)} - \tau(s_2). \tag{35}$$

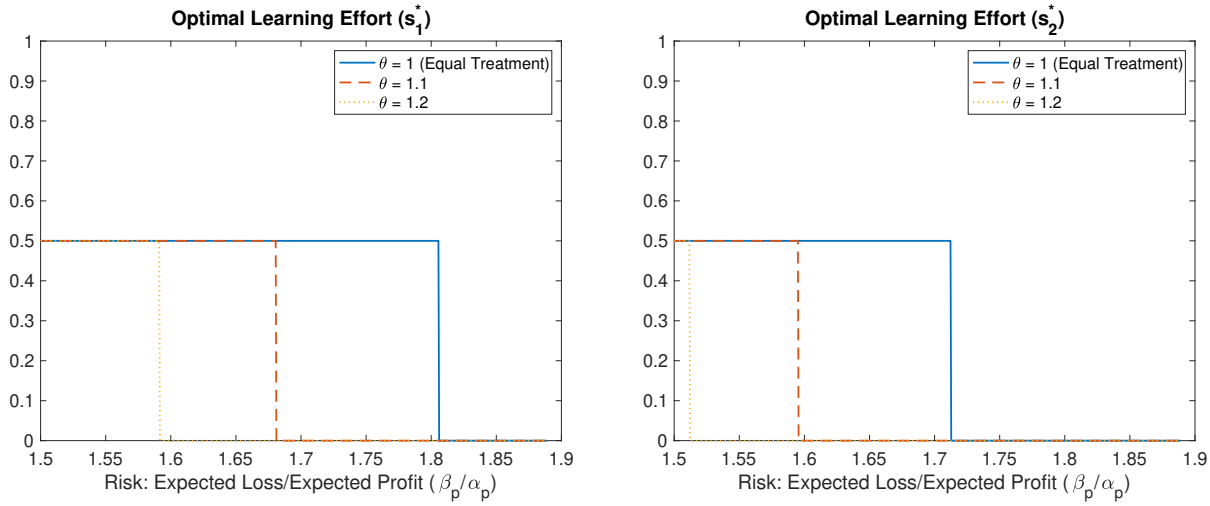

Figure 12 **Comparison of learning efforts: For all levels of risk in the market, (i) as θ increases, the optimal learning efforts in both the periods, i.e., s_1^* and s_2^* decrease. (ii) The optimal learning effort in the first period is greater than or equal to that in the second-period. Intuitively, this reflects the future value of the learning effort in the first period.**

The total profit is

$$\pi = \pi_1 + \pi_2.$$

We now numerically optimize the profit of the decision-maker and plot the relevant quantities of interest.

We can see that Figure 12 is qualitatively very similar to Figure 9. That is, as θ increases, the optimal learning efforts in both the periods, i.e., s_1^* and s_2^*, decrease. Also, the optimal learning effort in the first period is greater than or equal to that in the second-period. Intuitively, this reflects the future value of the learning effort in the first period. Similarly, Figure 13 and Figure 10 are also qualitatively similar. That is, in a risky market, fewer protected applicants are accepted when θ is high. Also, Figure 14 and Figure 11 reflect that the profit of decision maker decreases with θ. Overall, our main results continue to hold when the second-period learning rate depends on the number of accepted applicants in the first period.

We now proceed to analyzing a competitive setting to ensure the robustness of our main results under competition.

6.1. Competition

Our model thus far assumes a monopolist decision-maker. We now relax this assumption and consider a competitive setting. For tractability, we consider the model of Section 5 and assume that the

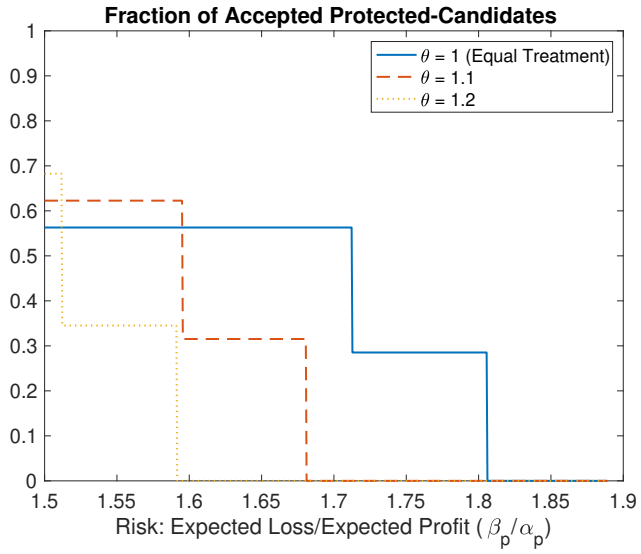

Figure 13 **In a risky market, fewer protected applicants are accepted as θ increases.**

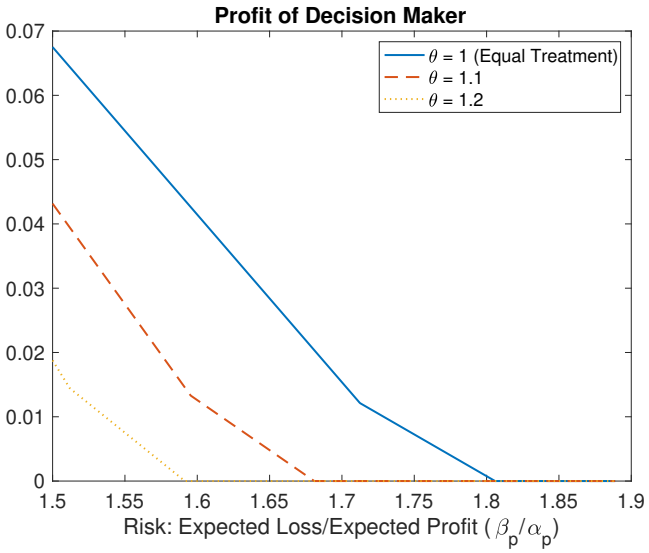

Figure 14 **Profit of the decision maker decreases with θ.**

decision-maker is operating under perfect competition. We assume that there are a large number of identical decision-makers competing on the revenue they receive from applicants (e.g., the interest charged by banks). In our model, this revenue is represented by α, because α is the utility of selecting a good applicant. As all the decision-makers are identical, we focus on a symmetric case in which each decision maker chooses the same value of α. One difference with the traditional models of competition is that the decision-maker can choose the cost of learning, i.e., $\tau(s)$, by deciding the learning effort s. Therefore, all combinations of α and s that lead to zero profit are equilibria. We

show that as θ increases, the learning effort exerted by the decision maker, i.e., s, decreases in all equilibria.

Using $\pi_\theta = 0$ in (13), we get

$$\pi_\theta = \frac{\gamma_r s \alpha_r}{1-\gamma_r s} + \frac{s G_1(\theta)}{1-\gamma_p s} - \tau(s) = 0. \tag{36}$$

Substituting $\tau(s) = \frac{ks}{1-\gamma_r s}$ and solving for s, we get

$$s^* = \frac{H_1(\theta) - 1}{H_1(\theta)\gamma_r - \gamma_p}, \tag{37}$$

where $H_1(\theta) = \frac{G_1(\theta)}{k - \gamma_r \alpha_r}$ and $G_1(\theta) = \gamma_p \beta_p - \theta \gamma_r (\beta_p - \alpha_p)$. Taking derivative of s^* with respect to θ we have

$$\frac{ds^*}{d\theta} = \frac{\gamma_r - \gamma_p}{[H_1(\theta)\gamma_r - \gamma_p]^2} \frac{dH_1(\theta)}{d\theta}. \tag{38}$$

It is easy to see that $\frac{ds^*}{d\theta} < 0$, because $\frac{dH_1(\theta)}{d\theta} < 0$. Formally, we have the following result.

Theorem 7 *The optimal learning effort $s^*(\theta)$ decreases with an increase in θ, under competition.*

The above theorem shows that the decision makers have lower incentive to exert effort in learning about applicants in a competitive setting. Intuitively, because learning is costly, a stringent fairness constraint under competition disincentivizes firms to invest in learning about the applicants.

6.2. Applicants' Response

Our base model does not consider applicants' strategic response to different fairness constraints and assumes that the fraction of bad applicants in the two groups are fixed. In reality, applicants may react to certain fairness constraints differently, e.g., make more efforts to become good applicants, and thus change the distribution of applicant types. In particular, Shimao et al. (2018) shows that EO-based machine learning equally incentivizes two demographic groups to put in efforts and improve skills, while other fairness constraints disproportionately discourage the protected groups from making efforts. In this section, we consider applicants' response and relax the assumption on the fixed fraction of bad applicants.

Specifically, we assume that under ET, the fraction of bad applicants (e.g., default rate) for the regular group and that for the protected group are still d_r and d_p, respectively, where $d_p > d_r$; under EO the fraction of bad applicants is identical for both groups at d_r. This reflects the idea that compared with ET, EO decreases the proportion of bad applicants as it equally incentivizes applicants from the two groups to put in effort and become good applicants. We further assume that the benefit of approving a good applicant (e.g., interest rate) would increase the overall default

rate increases to balance the risk. Specifically, we fix the expected loss rate of approving a random applicant:

$$\beta(\frac{d_p+d_r}{2})-\alpha(1-\frac{d_p+d_r}{2})=c \tag{39}$$

where c is a positive constant. This is equivalent to $(\beta_p+\beta_r)-(\alpha_p+\alpha_r)=2c$ and it specifies that

$$\alpha(d_p)=\frac{\beta(d_p+d_r)-2c}{2-d_p-d_r} \tag{40}$$

Previous analysis has shown that when the two groups have the same default rate (i.e., $d_r=d_p$), the optimal learning effort is higher under ET than under EO. We now proceed to analyze how the optimal learning effort under ET changes as d_p increases. Following the analysis in Section 4, we know that the profit under ET is:

$$\pi_{ET}=\frac{\alpha_r\gamma_r s}{1-\gamma_r s}+\frac{(\alpha_p\gamma_r+\beta_p\gamma_p-\beta_p\gamma_r)s}{1-\gamma_p s}-\tau(s) \tag{41}$$

Taking the derivative of the profit function with respect to s, we get

$$\frac{d\pi_{ET}}{ds}=\frac{\alpha_r\gamma_r}{(1-\gamma_r s)^2}+\frac{\alpha_p\gamma_r+\beta_p\gamma_p-\beta_p\gamma_r}{(1-\gamma_p s)^2}-\tau'(s) \tag{42}$$

$$=(\frac{1}{(1-\gamma_r s)^2}-\frac{1}{(1-\gamma_p s)^2})\alpha_r\gamma_r+\frac{(\alpha_r+\alpha_p)\gamma_r+\beta_p\gamma_p-\beta_p\gamma_r}{(1-\gamma_p s)^2}-\tau'(s) \tag{43}$$

As $\alpha_r+\alpha_p=\beta_r+\beta_p-2c$, we have:

$$\frac{d\pi_{ET}}{ds}=(\frac{1}{(1-\gamma_r s)^2}-\frac{1}{(1-\gamma_p s)^2})\alpha_r\gamma_r+\frac{(\beta_r+\beta_p-2c)\gamma_r+\beta_p\gamma_p-\beta_p\gamma_r}{(1-\gamma_p s)^2}-\tau'(s) \tag{44}$$

$$=(\frac{1}{(1-\gamma_r s)^2}-\frac{1}{(1-\gamma_p s)^2})\alpha_r\gamma_r+\frac{\beta_p\gamma_p+\beta_r\gamma_r-2c\gamma_r}{(1-\gamma_p s)^2}-\tau'(s) \tag{45}$$

As

$$\frac{d\alpha_r}{dd_p}=(1-d_r)\frac{d\alpha}{dd_p}>0,\quad \frac{d\beta_p}{dd_p}=\frac{d\beta d_p}{dd_p}=\beta>0, \tag{46}$$

it is easy to verify that $\frac{d}{dd_p}(\frac{d\pi_{ET}}{ds})>0$. Formally, we have the following result.

Theorem 8 *Optimal learning effort under ET increases with default rate in the protected group. That is, s^{ET} increases with d_p.*

Intuitively, when the default rate is higher, the decision-maker is motivated to learn more because the reward for separating good and bad applicants is larger. Combining this result with the previous result that the optimal learning is higher under ET than under EO when $d_p=d_r$, we have that the optimal learning effort is still higher under ET when we consider the change in default rates as a result of applicants' response.

7. Conclusion

Concerns about algorithmic bias have been raised since we realized that the seemingly innocuous machine learning algorithms could automate or even magnify human bias and inequity encoded in data. While trying our best to address these problems, we should be cautious about our approaches. Apart from the debate on the ultimate meaning of fairness and the worry about reverse discrimination, some well-intended and appealing fairness constraints may not necessarily help the protected group. When proposing the notion of equal opportunity, Hardt et al. (2016) argued that requiring equal opportunity threshold "transfers the burden of uncertainty from the protected class to the decision maker" and it "incentivizes the decision maker to invest additional resources toward building a better model." However, a better model does not come for free. In this paper, we show that when learning effort is an endogenous decision variable, the decision-maker would actually choose to invest fewer resources in building a good model under the constraint of equal opportunity compared with under the current legislation that requires equal treatment. This strategic behavior could reduce the decision-maker's loss, but would harm the candidates. When the market is risky, the harm of the reduced learning effort outweighs the direct benefit of a lower threshold, and everyone, including the very group that equal opportunity aims to protected, is worse off. More broadly, our results suggest that any fairness constraint that tries to close the gap in prediction outcomes by lowering the threshold for the protected would cause the same problem: reduced learning and potential harm to everyone.

This work is closely related to the literature on discussion of fairness notions. Kleinberg et al. (2016) and Chouldechova (2017) proved that several popular fairness notions cannot be satisfied simultaneously, and pointed out the inherent trade-offs among them. Liu et al. (2018) showed that equal opportunity and demographic parity sometimes may cause harm to the protected group in the long term due to the delayed impact of decisions. They used loan application as an example and considered the fact that a defaulted loan not only reduces the decision maker's profit, but also hurts the borrower's score. Even though the affirmative action of giving more loans to the protected group seems to be a blessing at the time of granting loans, it could eventually harm the score distribution in the group and put the protected candidates into a worse situation if many of the approved candidates are bad candidates. Kleinberg and Mullainathan (2018) showed that while we often favor a simple and interpretable model, a simplified model suffers from less accuracy and reduces utility for the disadvantaged (protected) group, and therefore creates inequity. Corbett-Davies and Goel (2018) criticized a family of fairness definitions called "classification parity," which includes equal opportunity and demographic parity. They argued that single threshold applied to the true risk distribution is optimal and it satisfies the compelling notion of equity that everyone is held

to the same standard, but it inevitably violates classification parity when the true risk distribution differs for two groups, which is almost always the case. While prior studies have pointed out the problems associated with the new fairness notions from different perspectives, this paper, to our best knowledge, is the first study that focuses on the dynamics of learning effort and considers decision makers' strategic behavior in response to fairness policies.

It is important to note that our paper does not intend to speak for the status quo of equal treatment requirement. Instead, we aim to highlight the importance of understanding the causes of algorithmic bias and considering strategic responses to fairness policies. Apparently our stylized model does not capture all the important elements in the decision-making process. For example, we do not consider the case where the firm is biased and gains utility from favoring the regular candidates, which is the primary reason for which we need anti-discrimination law. Nonetheless, the model illustrates that when the learning efficiency gap drives the discriminatory outcomes, forcefully closing the gap on the final decisions according to a simple statistic would bring unintended consequences and may even cause larger problems. We believe the more effective and efficient approaches to fairness requires addressing the source of biased outcomes, which in our case means reducing the learning efficiency gap. This requires efforts in a broader perspective than adjusting thresholds. With careful designs and regulations, algorithms could potentially improve both equity and efficacy, and therefore help use make better decisions.

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias. *ProPublica, May* 23.

Barocas, S., and A. Selbst. 2016. Big Data' s Disparate Impact. *California law review* 104 (1): 671–729.

Calders, T., F. Kamiran, and M. Pechenizkiy. 2009. Building Classifiers with Independency Constraints. *IEEE International Conference on Data Mining Workshops*:13–18.

Chandler, M. B. 1979. The Business Necessity Defense to Disparate-Impact Liability Under Title VII. *The University of Chicago Law Review* 46 (4): 911–934.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *ArXiv Preprint ArXiv ID: 1703.00056*.

Chouldechova, A., and M. G'Sell. 2017. Fairer and more accurate, but for whom? *arXiv preprint arXiv:1707.00046*.

Chouldechova, A., E. Putnam-Hornstein, D. Benavides-Prado, O. Fialko, and R. Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 81:134–148.

Chouldechova, A., and A. Roth. 2018. The Frontiers of Fairness in Machine Learning. *CoRR* abs/1810.08810.

Corbett-Davies, S., and S. Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *ArXiv Preprint ArXiv ID: 1808.00023*.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Fu, R., Y. Huang, and P. V. Singh. 2020. Crowd, Lending, Machine, and Bias. *arXiv preprint arXiv:2008.04068*.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. 2017. Predictably Unequal? The Effects of Machine Learning on Credit Markets.

Haghpanah, N., and R. Siegel. 2019. Pareto Improving Segmentation of Multi-product Markets. Technical report, Working Paper.

Hardt, M. 2014. How big data is unfair. `https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de`.

Hardt, M., E. Price, and N. Srebro. 2016, oct. Equality of Opportunity in Supervised Learning. *30th Conference on Neural Information Processing Systems*.

Kim, P. T. 2017. Data-Driven Discrimination at Work. *WILLIAM & MARY LAW REVIEW* 58:857–936.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2017. Human decisions and machine predictions. *The quarterly journal of economics* 133 (1): 237–293.

Kleinberg, J., and S. Mullainathan. 2018. Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability. *ArXiv Preprint ArXiv ID: 1809.04578*.

Kleinberg, J., S. Mullainathan, and M. Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *ArXiv Preprint ArXiv ID: 1609.05807*.

Liu, L. T., S. Dean, E. Rolf, M. Simchowitz, and M. Hardt. 2018. Delayed Impact of Fair Machine Learning. *ArXiv Preprint ArXiv ID: 1803.04383*.

Netzer, O., A. Lemaire, and M. Herzenstein. 2018. When words sweat: Identifying signals for loan default in the text of loan applications. *Columbia Business School Research Paper* (16-83).

ProPublica 2014. Machine Bias. Available at: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing` (Accessed May 1, 2019).

Shimao, H., J. Komiyama, and W. Khern-am nuai. 2018. The Cost of Fairness: Evaluating Economic Implications of Fairness-Aware Machine Learning. *Available at SSRN 3314373*.

Shimao, H., J. Komiyama, W. Khern-am nuai, and K. N. Kannan. 2019. Strategic Best-Response Fairness in Fair Machine Learning Algorithms. *Available at SSRN 3389631*.

Skeem, J. L., and C. T. Lowenkamp. 2016. Risk, Race, and Recidivism: Predictive Bias and Disparate Impact. *Criminology* 54 (4): 680–712.

Washington Post 2019. Apple Card algorithm sparks gender bias allegations against Goldman Sachs. Available at: `https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/` (Accessed August 20, 2020).

Appendix A Proofs of Results

Proof of Theorem 1: To prove Theorem 1, we first prove the following statement: Let $f(x)$ and $g(x)$ be two unimodal continuous and differentiable functions on a closed interval $[a, b]$, and x_f^* and x_g^* be the x values that maximize $f(x)$ and $g(x)$, respectively. If $f'(x) \geq g'(x)$ for all x in the domain, then $x_f^* \geq x_g^*$.

Because $f'(x) \geq g'(x)$, we have $f'(x_g^*) \geq g'(x_g^*)$. It means that the value of $f(x)$ can increase by choosing the value of x above x_g^*. Thus, there exists a better feasible solution for optimizing $f(x)$, and that feasible solution is higher than x_g^*. Since $f(x)$ is a unimodal function, we have $x_f^* \geq x_g^*$.

We now proceed to proving the main result in Theorem 1. The derivative of the two profit functions are:

$$\pi'_{\text{ET}}(s) = \frac{\alpha_p\gamma_r}{(1-\gamma_p s)^2} + \frac{\alpha_r\gamma_r}{(1-\gamma_r s)^2} + \frac{\beta_p\gamma_p}{(1-\gamma_p s)^2} - \frac{\beta_p\gamma_r}{(1-\gamma_p s)^2} - \tau'(s), \quad s \in [0, \frac{1}{2\gamma_r}];$$
$$\pi'_{\text{EO}}(s) = \frac{\alpha_p\gamma_r}{(1-\gamma_r s)^2} + \frac{\alpha_r\gamma_r}{(1-\gamma_r s)^2} + \frac{\beta_p\gamma_p}{(1-\gamma_p s)^2} - \frac{\beta_p\gamma_r}{(1-\gamma_r s)^2} - \tau'(s), \quad s \in [0, \frac{1}{2\gamma_r}].$$

It is easy to verify that for $\tau(s) = \frac{ks}{1-\gamma_r s}$, $\pi_{\text{ET}}(s)$ and $\pi_{\text{EO}}(s)$ are both unimodal functions. Using above equations, $\forall s \in [0, \frac{1}{2\gamma_r}]$ we have

$$\begin{aligned}\pi'_{\text{ET}}(s) - \pi'_{\text{EO}}(s) &= \big(\frac{\alpha_p\gamma_r}{(1-\gamma_p s)^2} - \frac{\beta_p\gamma_r}{(1-\gamma_p s)^2}\big) - \big(\frac{\alpha_p\gamma_r}{(1-\gamma_r s)^2} - \frac{\beta_p\gamma_r}{(1-\gamma_r s)^2}\big), \\ &= \gamma_r(\beta_p - \alpha_p)\big(\frac{1}{(1-\gamma_r s)^2} - \frac{1}{(1-\gamma_p s)^2}\big) \geq 0.\end{aligned}$$

Using the statement proven in the beginning of this proof, we have $s^{\text{ET}} \geq s^{\text{EO}}$. ■

Proof of Proposition 2: For any given $s \in [0, \frac{1}{2\gamma_r}]$, we have

$$\begin{aligned}\pi_{\text{ET}}(s) - \pi_{\text{EO}}(s) &= \big(\frac{\alpha_p\gamma_r s}{1-\gamma_p s} - \frac{\beta_p\gamma_r s}{1-\gamma_p s}\big) - \big(\frac{\alpha_p\gamma_r s}{1-\gamma_r s} - \frac{\beta_p\gamma_r}{1-\gamma_r s}\big) \\ &= \gamma_r(\beta_p - \alpha_p)\big(\frac{s}{1-\gamma_r s} - \frac{s}{1-\gamma_p s}\big) \geq 0.\end{aligned}$$

Therefore,

$$\pi_{\text{ET}}(s^{\text{EO}}) \geq \pi_{\text{EO}}(s^{\text{EO}}).$$

Since s^{ET} is the optimal amount of learning effort under ET,

$$\pi_{\text{ET}}(s^{\text{ET}}) \geq \pi_{\text{ET}}(s^{\text{EO}}).$$

By transitivity,

$$\pi_{\text{ET}}(s^{\text{ET}}) \geq \pi_{\text{EO}}(s^{\text{EO}}),$$

i.e., $\pi_{\text{ET}}^* \geq \pi_{\text{EO}}^*$. ■

Proof of Proposition 3:

$$\phi_r^{ET} = \frac{1-c_r^{\mathrm{ET}}}{1-\gamma_r s^{\mathrm{ET}}} = \frac{\gamma_r s^{\mathrm{ET}}}{1-\gamma_r s^{\mathrm{ET}}},$$
$$\phi_r^{EO} = \frac{1-c_r^{\mathrm{EO}}}{1-\gamma_r s^{\mathrm{EO}}} = \frac{\gamma_r s^{\mathrm{EO}}}{1-\gamma_r s^{\mathrm{EO}}}.$$

Theorem 1 shows

$$s^{\mathrm{ET}} \geq s^{\mathrm{EO}},$$

therefore,

$$\frac{\gamma_r s^{\mathrm{ET}}}{1-\gamma_r s^{\mathrm{ET}}} \geq \frac{\gamma_r s^{\mathrm{EO}}}{1-\gamma_r s^{\mathrm{EO}}},$$

i.e, $\phi_r^{ET} \geq \phi_r^{EO}$. ■

Proof of Theorem 2: First Order Condition gives us:

$$\pi'_{\mathrm{ET}}(s^{\mathrm{ET}}) = \frac{\alpha_p\gamma_r}{(1-\gamma_p s^{\mathrm{ET}})^2} + \frac{\alpha_r\gamma_r}{(1-\gamma_r s^{\mathrm{ET}})^2} + \frac{\beta_p\gamma_p}{(1-\gamma_p s^{\mathrm{ET}})^2} - \frac{\beta_p\gamma_r}{(1-\gamma_p s^{\mathrm{ET}})^2} - \frac{k}{(1-\gamma_r s^{\mathrm{ET}})^2} = 0; \quad (47)$$
$$\pi'_{\mathrm{EO}}(s^{\mathrm{EO}}) = \frac{\alpha_p\gamma_r}{(1-\gamma_r s^{\mathrm{EO}})^2} + \frac{\alpha_r\gamma_r}{(1-\gamma_r s^{\mathrm{EO}})^2} + \frac{\beta_p\gamma_p}{(1-\gamma_p s^{\mathrm{EO}})^2} - \frac{\beta_p\gamma_r}{(1-\gamma_r s^{\mathrm{EO}})^2} - \frac{k}{(1-\gamma_r s^{\mathrm{EO}})^2} = 0. \quad (48)$$

Multiply (47) by $(1-\gamma_r s^{\mathrm{ET}})^2$ and arrange the equation:

$$(\frac{1-\gamma_r s^{\mathrm{ET}}}{1-\gamma_p s^{\mathrm{ET}}})^2 \cdot (\alpha_p\gamma_r + \beta_p\gamma_p - \beta_p\gamma_r) = k - \alpha_r\gamma_r \quad (49)$$

Note that

$$\frac{1-\gamma_r s^{\mathrm{ET}}}{1-\gamma_p s^{\mathrm{ET}}} = 1 - (\gamma_r - \gamma_p)\frac{s^{\mathrm{ET}}}{1-\gamma_p s^{\mathrm{ET}}}, \quad (50)$$

Substituting (50) into (49), with some algebra we have

$$\phi_p^{\mathrm{ET}} \equiv \frac{\gamma_r s^{\mathrm{ET}}}{1-\gamma_p s^{\mathrm{ET}}} = \frac{\gamma_r}{\gamma_r - \gamma_p}[1 - (\frac{k-\alpha_r\gamma_r}{\alpha_p\gamma_r + \beta_p\gamma_p - \beta_p\gamma_r})^{\frac{1}{2}}] \quad (51)$$

Similarly, from (48) we derive the expression for the coverage rate under equal opportunity:

$$\phi_p^{\mathrm{EO}} \equiv \frac{\gamma_r s^{\mathrm{EO}}}{1-\gamma_r s^{\mathrm{EO}}} = \frac{\gamma_r}{\gamma_r - \gamma_p}[(\frac{\beta_p\gamma_p}{k-\alpha_r\gamma_r + \beta_p\gamma_r - \alpha_p\gamma_r})^{\frac{1}{2}} - 1] \quad (52)$$

Therefore, we have:

$$\phi_p^{\mathrm{ET}} - \phi_p^{\mathrm{EO}} = \frac{\gamma_r}{\gamma_r - \gamma_p}[2 - (\frac{k-\alpha_r\gamma_r}{\alpha_p\gamma_r + \beta_p\gamma_p - \beta_p\gamma_r})^{\frac{1}{2}} - (\frac{\beta_p\gamma_p}{k-\alpha_r\gamma_r + \beta_p\gamma_r - \alpha_p\gamma_r})^{\frac{1}{2}}] \quad (53)$$
$$= \frac{1}{\gamma_r - \gamma_p}[2 - \sigma - (\frac{\beta_p\gamma_p}{\sigma^2\beta_p\gamma_p + (1-\sigma^2)(\beta_p\gamma_r - \alpha_p\gamma_r)})^{\frac{1}{2}}] \quad (54)$$

When $\frac{\alpha_p}{\beta_p} \leq 1 - \frac{1-(2-\sigma)^2)\sigma^2}{(1-\sigma^2)(2-\sigma)^2} \cdot \frac{\gamma_p}{\gamma_r}$, we have

$$(\beta_p - \alpha_p)\gamma_r \geq \frac{1-(2-\sigma)^2)\sigma^2}{(1-\sigma^2)(2-\sigma)^2} \cdot \beta_p\gamma_p, \tag{55}$$

$$[1-(2-\sigma)^2\sigma^2]\beta_p\gamma_p \leq (1-\sigma^2)(2-\sigma)^2(\beta_p\gamma_r - \alpha_p\gamma_r), \tag{56}$$

$$\beta_p\gamma_p \leq (2-\sigma)^2\sigma^2\beta_p\gamma_p + (2-\sigma)^2(1-\sigma^2)(\beta_p\gamma_r - \alpha_p\gamma_r), \tag{57}$$

$$\frac{\beta_p\gamma_p}{\sigma^2\beta_p\gamma_p + (1-\sigma^2)(\beta_p\gamma_r - \alpha_p\gamma_r)} \leq (2-\sigma)^2, \tag{58}$$

therefore,

$$(\frac{\beta_p\gamma_p}{\sigma^2\beta_p\gamma_p + (1-\sigma^2)(\beta_p\gamma_r - \alpha_p\gamma_r)})^{\frac{1}{2}} \leq 2-\sigma, \tag{59}$$

which implies that $\phi_p^{\mathrm{ET}} > \phi_p^{\mathrm{EO}}$, since $\gamma_r - \gamma_p > 0$. ■

Proof of Theorem 3: Let NS_p^{ET} and NF_p^{ET} be the number of successful acceptance and the number of failed acceptance in the protected group under equal treatment, respectively. Then,

$$NS_p^{\mathrm{ET}} = \frac{1-c^{\mathrm{ET}}}{1-\gamma_p s^{\mathrm{ET}}}(1-d_p) = \frac{\gamma_r s^{\mathrm{ET}}}{1-\gamma_p s^{\mathrm{ET}}}(1-d_p);$$
$$NF_p^{\mathrm{ET}} = \frac{1-\gamma_p s^{\mathrm{ET}} - c^{\mathrm{ET}}}{1-\gamma_p s^{\mathrm{ET}}} d_p = \frac{(\gamma_r - \gamma_p)s^{\mathrm{ET}}}{1-\gamma_p s^{\mathrm{ET}}} d.$$

Therefore, we have

$$\delta_p^{\mathrm{ET}} = \frac{NS_p^{\mathrm{ET}}}{NS_p^{\mathrm{ET}} + NF_p^{\mathrm{ET}}} = \frac{(1-d_p)\gamma_r}{\gamma_r - \gamma_p d_p}.$$

Similarly, under equal opportunity, we have

$$NS_p^{\mathrm{EO}} = \frac{1-c_p^{\mathrm{EO}}}{1-\gamma_p s^{\mathrm{EO}}}(1-d_p) = \frac{\gamma_r s^{\mathrm{EO}}}{1-\gamma_r s^{\mathrm{EO}}}(1-d_p)$$
$$NF_p^{\mathrm{EO}} = \frac{1-\gamma_p s^{\mathrm{EO}} - c_p^{\mathrm{EO}}}{1-\gamma_p s^{\mathrm{EO}}} d_p = (\frac{\gamma_r s^{\mathrm{EO}}}{1-\gamma_r s^{\mathrm{EO}}} - \frac{\gamma_p s^{\mathrm{EO}}}{1-\gamma_p s^{\mathrm{EO}}}) \cdot d_p$$

Thus,

$$\delta_p^{\mathrm{EO}} = \frac{NS_p^{\mathrm{EO}}}{NS_p^{\mathrm{ET}} + NF_p^{\mathrm{ET}}} = \frac{(1-d_p)\gamma_r}{\gamma_r - \frac{1-\gamma_r s^{\mathrm{EO}}}{1-\gamma_p s^{\mathrm{EO}}} \cdot \gamma_p d_p} < \frac{(1-d_p)\gamma_r}{\gamma_r - \gamma_p d_p} = \delta_p^{\mathrm{ET}}.$$

■

Appendix B Proofs of the Results in the Two Other Cases

B.1 Medium Expected Loss

In this section, we show that our results hold for the case of medium expected loss, i.e.,

$$\alpha_p + \alpha_r < \beta_p \leq \alpha_p + \frac{1-\gamma_p s}{1-\gamma_r s}\alpha_r.$$

From the analysis in section 3, we know that in this case

$$c_p^{\mathrm{ET}} = c_r^{\mathrm{ET}} = 1-\gamma_r s, \qquad \pi_{\mathrm{ET}}(s) = \frac{\alpha_p\gamma_r s}{1-\gamma_p s} + \frac{\alpha_r\gamma_r s}{1-\gamma_r s} + \frac{\beta_p\gamma_p s}{1-\gamma_p s} - \frac{\beta_p\gamma_r s}{1-\gamma_p s} - \tau(s)$$
$$c_p^{\mathrm{EO}} = 1-\gamma_p s, c_r^{\mathrm{EO}} = 1-\frac{1-\gamma_r s}{1-\gamma_p s}\gamma_p s, \qquad \pi_{\mathrm{EO}}(s) = \frac{\alpha_p\gamma_p s}{1-\gamma_p s} + \frac{\alpha_r\gamma_p s}{1-\gamma_p s} - \tau(s);$$

Learning

The derivative of the two profit functions are:

$$\pi'_{\mathrm{ET}}(s) = \frac{\alpha_p\gamma_r}{(1-\gamma_p s)^2} + \frac{\alpha_r\gamma_r}{(1-\gamma_r s)^2} + \frac{\beta_p\gamma_p}{(1-\gamma_p s)^2} - \frac{\beta_p\gamma_r}{(1-\gamma_p s)^2} - \tau'(s)$$
$$\pi'_{\mathrm{EO}}(s) = \frac{\alpha_p\gamma_p}{(1-\gamma_p s)^2} + \frac{\alpha_r\gamma_p}{(1-\gamma_p s)^2} - \tau'(s);$$

Therefore,

$$\begin{aligned}\pi'_{\mathrm{ET}}(s) - \pi'_{\mathrm{EO}}(s) &= \frac{\alpha_p\gamma_r}{(1-\gamma_p s)^2} + \frac{\alpha_r\gamma_r}{(1-\gamma_r s)^2} + \frac{\beta_p\gamma_p}{(1-\gamma_p s)^2} - \frac{\beta_p\gamma_r}{(1-\gamma_p s)^2} - \frac{\alpha_p\gamma_p}{(1-\gamma_p s)^2} - \frac{\alpha_r\gamma_p}{(1-\gamma_p s)^2} \\ &= \frac{(\beta_p-\alpha_p-\alpha_r)\gamma_p}{(1-\gamma_p s)^2} + [\frac{\alpha_r}{(1-\gamma_r s)^2} - \frac{\beta_p}{(1-\gamma_p s)^2}]\gamma_r\end{aligned}$$

As

$$\beta_p \le \alpha_p + \frac{1-\gamma_p s}{1-\gamma_r s}\alpha_r,$$

we have

$$\beta_p < \frac{1-\gamma_p s}{1-\gamma_r s}\alpha_r < (\frac{1-\gamma_p s}{1-\gamma_r s})^2\alpha_r,$$

thus,

$$\frac{\beta_p}{(1-\gamma_p s)^2} < \frac{\alpha_r}{(1-\gamma_r s)^2}.$$

Also

$$\beta_p > \alpha_p + \alpha_r.$$

Therefore,

$$\pi'_{\mathrm{ET}}(s) - \pi'_{\mathrm{EO}}(s) > 0.$$

By the statement shown in the proof of Theorem 1, we have $s^{\mathrm{ET}} \ge s^{\mathrm{EO}}$. ■

Impact on the decision-maker

For any given $s \in [0, \frac{1}{2\gamma_r}]$, we have

$$\pi_{\mathrm{ET}}(s) - \pi_{\mathrm{EO}}(s) = \frac{(\beta_p-\alpha_p-\alpha_r)\gamma_p s}{1-\gamma_p s} + [\frac{\alpha_r}{1-\gamma_r s} - \frac{\beta_p}{1-\gamma_p s}]\gamma_r s \ge 0$$

Therefore,

$$\pi_{\mathrm{ET}}(s^{\mathrm{EO}}) \geq \pi_{\mathrm{EO}}(s^{\mathrm{EO}}).$$

Since s^{ET} is the optimal amount of learning effort under ET,

$$\pi_{\mathrm{ET}}(s^{\mathrm{ET}}) \geq \pi_{\mathrm{ET}}(s^{\mathrm{EO}}).$$

By transitivity,

$$\pi_{\mathrm{ET}}(s^{\mathrm{ET}}) \geq \pi_{\mathrm{EO}}(s^{\mathrm{EO}}),$$

i.e., $\pi^*_{\mathrm{ET}} \geq \pi^*_{\mathrm{EO}}$. ■

Impact on the Regular Group

$$\phi_r^{ET} = \frac{1 - c_r^{\mathrm{ET}}}{1 - \gamma_r s^{\mathrm{ET}}} = \frac{\gamma_r s^{\mathrm{ET}}}{1 - \gamma_r s^{\mathrm{ET}}},$$
$$\phi_r^{EO} = \frac{1 - c_r^{\mathrm{EO}}}{1 - \gamma_r s^{\mathrm{EO}}} = \frac{\gamma_p s^{\mathrm{EO}}}{1 - \gamma_r s^{\mathrm{EO}}}.$$

Since

$$s^{\mathrm{ET}} \geq s^{\mathrm{EO}}, \gamma_r \geq \gamma_p$$

therefore,

$$\frac{\gamma_r s^{\mathrm{ET}}}{1 - \gamma_r s^{\mathrm{ET}}} \geq \frac{\gamma_r s^{\mathrm{EO}}}{1 - \gamma_r s^{\mathrm{EO}}} \geq \frac{\gamma_p s^{\mathrm{EO}}}{1 - \gamma_r s^{\mathrm{EO}}},$$

i.e, $\phi_r^{ET} \geq \phi_r^{EO}$.

Since $c_p^{\mathrm{ET}}, c_p^{\mathrm{EO}} \geq 1 - \gamma_r s$, we have $\delta_r^{\mathrm{ET}} = \delta_r^{\mathrm{EO}} = 1$. ■

Impact on the Protected Group

$$\phi_p^{ET} = \frac{1 - c_p^{\mathrm{ET}}}{1 - \gamma_p s^{\mathrm{ET}}} = \frac{\gamma_r s^{\mathrm{ET}}}{1 - \gamma_p s^{\mathrm{ET}}},$$
$$\phi_r^{EO} = \frac{1 - c_p^{\mathrm{EO}}}{1 - \gamma_p s^{\mathrm{EO}}} = \frac{\gamma_p s^{\mathrm{EO}}}{1 - \gamma_p s^{\mathrm{EO}}}.$$

Since

$$s^{\mathrm{ET}} \geq s^{\mathrm{EO}}, \gamma_r \geq \gamma_p,$$

we have $\phi_p^{ET} \geq \phi_r^{EO}$. ■

B.2 Large Expected Loss

In this section, we show that our results hold for the case of large expected loss, i.e.,

$$\beta_p > \alpha_p + \frac{1-\gamma_p s}{1-\gamma_r s}\alpha_r.$$

From the analysis in section 3, we know that in this case

$$c_p^{\mathrm{ET}} = c_r^{\mathrm{ET}} = 1-\gamma_p s, \qquad \pi_{\mathrm{ET}}(s) = \frac{\alpha_p\gamma_p s}{1-\gamma_p s} + \frac{\alpha_r\gamma_p s}{1-\gamma_r s} - \tau(s);$$
$$c_p^{\mathrm{EO}} = 1-\gamma_p s, c_r^{\mathrm{EO}} = 1 - \frac{1-\gamma_r s}{1-\gamma_p s}\gamma_p s, \qquad \pi_{\mathrm{EO}}(s) = \frac{\alpha_p\gamma_p s}{1-\gamma_p s} + \frac{\alpha_r\gamma_p s}{1-\gamma_p s} - \tau(s).$$

Learning

The derivative of the two profit functions are:

$$\pi'_{\mathrm{ET}}(s) = \frac{\alpha_p\gamma_p}{(1-\gamma_p s)^2} + \frac{\alpha_r\gamma_p}{(1-\gamma_r s)^2} - \tau'(s);$$
$$\pi'_{\mathrm{EO}}(s) = \frac{\alpha_p\gamma_p}{(1-\gamma_p s)^2} + \frac{\alpha_r\gamma_p}{(1-\gamma_p s)^2} - \tau'(s).$$

Therefore,

$$\pi'_{\mathrm{ET}}(s) - \pi'_{\mathrm{EO}}(s) = \frac{\alpha_r\gamma_p}{(1-\gamma_r s)^2} - \frac{\alpha_r\gamma_p}{(1-\gamma_p s)^2} \geq 0.$$

By the statement shown in the proof of Theorem 1, we have $s^{\mathrm{ET}} \geq s^{\mathrm{EO}}$. ■

Impact on the decision maker

For any given $s \in [0, \frac{1}{2\gamma_r}]$, we have

$$\pi_{\mathrm{ET}}(s) - \pi_{\mathrm{EO}}(s) = \frac{\alpha_r\gamma_p s}{1-\gamma_r s} - \frac{\alpha_r\gamma_p s}{1-\gamma_p s} \geq 0$$

Therefore,

$$\pi_{\mathrm{ET}}(s^{\mathrm{EO}}) \geq \pi_{\mathrm{EO}}(s^{\mathrm{EO}}).$$

Since s^{ET} is the optimal amount of learning effort under ET,

$$\pi_{\mathrm{ET}}(s^{\mathrm{ET}}) \geq \pi_{\mathrm{ET}}(s^{\mathrm{EO}}).$$

By transitivity,

$$\pi_{\mathrm{ET}}(s^{\mathrm{ET}}) \geq \pi_{\mathrm{EO}}(s^{\mathrm{EO}}),$$

i.e., $\pi^*_{\mathrm{ET}} \geq \pi^*_{\mathrm{EO}}$. ■

Impact on the Regular Group

$$\phi_r^{ET} = \frac{1 - c_r^{\mathrm{ET}}}{1 - \gamma_r s^{\mathrm{ET}}} = \frac{\gamma_p s^{\mathrm{ET}}}{1 - \gamma_r s^{\mathrm{ET}}},$$
$$\phi_r^{EO} = \frac{1 - c_r^{\mathrm{EO}}}{1 - \gamma_r s^{\mathrm{EO}}} = \frac{\gamma_p s^{\mathrm{EO}}}{1 - \gamma_p s^{\mathrm{EO}}}.$$

Since

$$s^{\mathrm{ET}} \geq s^{\mathrm{EO}}, \gamma_r \geq \gamma_p$$

therefore,

$$\frac{\gamma_p s^{\mathrm{ET}}}{1 - \gamma_r s^{\mathrm{ET}}} \geq \frac{\gamma_p s^{\mathrm{ET}}}{1 - \gamma_p s^{\mathrm{ET}}} \geq \frac{\gamma_p s^{\mathrm{EO}}}{1 - \gamma_p s^{\mathrm{EO}}},$$

i.e, $\phi_r^{ET} \geq \phi_r^{EO}$.

As $c_p^{\mathrm{ET}}, c_p^{\mathrm{EO}} \geq 1 - \gamma_r s$, we have $\delta_r^{\mathrm{ET}} = \delta_r^{\mathrm{EO}} = 1$. ■

Impact on the Protected Group

$$\phi_p^{ET} = \frac{1 - c_p^{\mathrm{ET}}}{1 - \gamma_p s^{\mathrm{ET}}} = \frac{\gamma_p s^{\mathrm{ET}}}{1 - \gamma_p s^{\mathrm{ET}}},$$
$$\phi_r^{EO} = \frac{1 - c_p^{\mathrm{EO}}}{1 - \gamma_p s^{\mathrm{EO}}} = \frac{\gamma_p s^{\mathrm{EO}}}{1 - \gamma_p s^{\mathrm{EO}}}.$$

Since

$$s^{\mathrm{ET}} \geq s^{\mathrm{EO}}$$

we have $\phi_p^{ET} \geq \phi_r^{EO}$. ■