PRODUCTION AND OPERATIONS MANAGEMENT
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1059-1478 | EISSN 1937-5956 | 00 | 0000 | 0001

POMS
DOI 10.1287/xxxx.0000.0000
© 0000 POMS

Goal Orientation for Fair Machine Learning Algorithms

Heng Xu
Warrington College of Business, University of Florida, heng.xu@ufl.edu

Nan Zhang
Warrington College of Business, University of Florida, zhang.nan@ufl.edu

Abstract: A key challenge facing the use of Machine Learning (ML) in organizational selection settings (e.g., the processing of loan or job applications) is the potential bias against (racial and gender) minorities. To address this challenge, a rich literature of Fairness-Aware ML (FAML) algorithms has emerged, attempting to ameliorate biases while maintaining the predictive accuracy of ML algorithms. Almost all existing FAML algorithms define their optimization goals according to a *selection task*, meaning that ML outputs are assumed to be the final selection outcome. In practice, though, ML outputs are rarely used as-is. In personnel selection, for example, ML often serves a support role to human resource managers, allowing them to more easily exclude unqualified applicants. This effectively assigns to ML a *screening* rather than selection task. It might be tempting to treat selection and screening as two variations of the same task that differ only quantitatively on the admission rate. This paper, however, reveals a qualitative difference between the two in terms of fairness. Specifically, we demonstrate through conceptual development and mathematical analysis that mis-categorizing a screening task as a selection one could not only degrade final selection quality but result in fairness problems such as selection biases within the minority group. After validating our findings with experimental studies on simulated and real-world data, we discuss several business and policy implications, highlighting the need for firms and policymakers to properly categorize the task assigned to ML in assessing and correcting algorithmic biases.

Key words: fairness, machine learning, optimization goal, selection, screening

1 Introduction

The past decade witnessed remarkable advances in the use of Machine Learning (ML) in operational selection processes such as the processing of loan or job applications (Mithas et al. 2022). In personnel selection, for example, ML is reportedly used in about one third of all organizations (Gonzalez et al. 2019). A particular appeal of using ML in these selection settings is the ease of casting the problem as predicting the *quality* of a selection outcome, e.g., the future job performance of applicants being selected, based on predictors such as the biodata and test scores of applicants. Once a firm collects historic data for these predictors and quality outcomes (e.g., from current/past employees), it runs an ML algorithm over the historic data to train a *prediction model*, before using the model in support of future selections.

Yet the use of ML in selection also faces an enormous challenge in terms of fairness across demographic groups (Sunar and Swaminathan 2022), such as those defined by legally protected characteristics including

What we submit in our current work, however, is that the two tasks differ *qualitatively* for the design of an FAML algorithm. As elaborated in the paper, a root distinction between the two is the cost/benefit tradeoff for FAML to make risky choices. Consider personnel selection as an example. Suppose that FAML predicts the quality (e.g., future job performance) of an applicant to follow a bimodal distribution[2]

background (e.g., attendance in women's colleges), which became a proxy for the protected gender variable.

we assume the training dataset to be sufficiently large, rendering the choice of technical design unimportant for conceptual/theoretical development in the paper.

Whereas the FAML literature now includes many algorithms that can satisfy both the ban on disparate treatment and the various types of fairness constraints over disparate impact (Mehrabi et al. 2021), researchers have also identified many concerns over the existing FAML algorithms, from a decrease of selection quality (Kleinberg et al. 2017) to the emergence of perverse incentives (Lipton et al. 2018), to sometimes exacerbating rather than ameliorating the bias in ML predictions (Corbett-Davies et al. 2017). Lipton et al. (2018), for example, note that these algorithms could create fairness issues *within* the minority group, basing their selections not on the predicted quality of a candidate but on whether the candidate "looks like" a minority according to the predictors.

To address these concerns, there were recent calls for abandoning the ban on disparate treatment (e.g., Lipton et al. 2018), instead legalizing an "algorithmic affirmative action" (Bent 2020). Doing so would allow the ML algorithm to become a "*decoupled classifier*" (Dwork et al. 2018), which assigns a separate quota to the minority and majority candidates, before learning separate prediction models for each group, so as to eliminate any within-group fairness issues. While the legal issues related to affirmative action are undoubtedly complex (Sackett and Wilk 1994), what we will submit in this paper is that there may be other ways to address the existing concerns on FAML *without* changing the law, e.g., by precisely defining the task assigned to ML in practice as a screening task rather than (over)simplifying it as a selection one.

3 Selection vs. Screening without Fairness Constraint

In this section, we examine the differences between ML for selection and screening *without* fairness conjob performance of the candidate, which cannot be observed but only predicted. With these notations, we can then summarize the population of candidates as a joint distribution G over the random vector $\langle \mathbf{x}; v; y \rangle$.

ML Selection Decisions: As discussed in Section 2.1.1, an ML algorithm is prohibited by law from accessing the group label (i.e., v) of a candidate. Since access to v is barred whereas y is unobservable, a selection decision made by ML can depend only on the characteristics $\mathbf{x}$ of a candidate. We therefore denote the ML *selection decision*

With this, the ML algorithm design is then reduced to generating an accurate point estimate of $\mathbb{E}_{\mathrm{G}}(yj\mathbf{x})$ for a given $\mathbf{x}$. To do so, the ML algorithm learns from a training dataset formed by historic instances of $h\mathbf{x};v;yi$ which are assumed to be drawn from the same joint distribution G. For example, in personnel selection, firms often train ML algorithms with data from incumbent (i.e., current and past) employees, using their past job applicants to populate $\mathbf{x}$, their demographic data to fill v, and their performance ratings (e.g., items scanned per minute for a supermarket checkout clerk, supervisor-rated performance, etc.) as y (Zhang et al. 2023). Unlike in the case of making predictions and selection decisions for candidates, where the ML algorithm cannot access v (legally) or y (practically), there is neither legal nor practical limit on what information the ML algorithm may learn from incumbent employees. Since the purpose of this paper is to examine the goal orientation for ML algorithms rather than the design of their learning processes, we assume the training dataset $h\mathbf{x};v;yi$ to be sufficiently large so as to allow ML to learn the joint distribution G to an arbitrary precision. We will relax this assumption later in experimental studies that use a real-world dataset.

3.2 ML for Screening Task

ML Screening and Manual Interviews: For the screening setting, we follow the exact same notations as in the selection setting. That is, when making screening decision for a candidate $h\mathbf{x};v;yi$ drawn from joint distribution G, the ML algorithm only has access to the candidate's characteristics vector $\mathbf{x}$, and admits the candidate with probability $L(\mathbf{x})\ 2\ [0;1]$. Unlike in the selection setting, the candidates admitted by the ML

i(2)3TJ/F3 11.7732423 s(a24omodel.9(a24odo)-240(a24othdrusings7732423ction242ics7(e28e,(ge23)w55 -18.4

orientationdatad339(a56ssu38(357)scter)10356 s

$$= \int_{W} \mathbb{E}_G(y \cdot \mathbb{1}(y \geq y_0) | \mathbf{x}) \cdot$$

distribution is bimodal, representing a high-risk high-reward choice. That is, admitting Bob could lead to a high reward (in terms of final selection quality) if he happens to be in the right component. Yet the decision is also risky because of the possibility for Bob to fall under the left, low-quality, component.

Now consider whether either algorithm prefers Alice or Bob in their output. As depicted in Figure 1a, Alice has a higher expected quality $\mathbb{E}_{\mathrm{G}}(y|\mathbf{x})$ than Bob, meaning that the selection algorithm would prefer Alice over Bob. In contrast, Figure 1b shows that, if we compare not the expected quality but the conditional expectation of quality given a positive interview outcome (i.e., $\mathbb{E}_{\mathrm{G}}(y|y \geq y_0, \mathbf{x})$), say with $y_0 = 4.5$, then Bob would have a higher expectation than Alice, meaning that the screening algorithm would prefer Bob over Alice. The root reason for this difference, as depicted in Figure 1b, is that the manual interview *de-risks* the selection of Bob. That is, if Bob happens to be in the left (i.e., low-quality) component, he will

where $\mathbb{T}$, as defined in Section 3.2, is the binary outcome indicator for the manual interview. Using the same simplification of low interview cost in Section 3.2 and the same method of Lagrange multiplier as Equation 12, we can simplify Equation 14 to

$$\begin{aligned} L^* &= \arg\max_{L} \int_{\Omega} (\mathbb{E}_G(y \cdot \mathbb{T}|\mathbf{x}) + \lambda \cdot \Pr\{v = 1, \mathbb{T} = 1|\mathbf{x}\}) \cdot L(\mathbf{x}) \cdot p_G(\mathbf{x}) d\mathbf{x} \\ &= \arg\max_{L} \int_{\Omega} (\mathbb{E}_G(y|\mathbb{T} = 1, \mathbf{x}) \cdot \Pr\{\mathbb{T} = 1|\mathbf{x}\} + \lambda \cdot \Pr\{v = 1|\mathbb{T} = 1, \mathbf{x}\} \cdot \Pr\{\mathbb{T} = 1|\mathbf{x}\}) \cdot L(\mathbf{x}) \cdot p_G(\mathbf{x}) d\mathbf{x} \\ &= \arg\max_{L} \int_{\Omega} (\mathbb{E}_G(y|\mathbb{T} = 1, \mathbf{x}) + \lambda \cdot \Pr\{v = 1|\mathbb{T} = 1, \mathbf{x}\}) \cdot \Pr\{\mathbb{T} = 1|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_G(\mathbf{x}) d\mathbf{x} \\ s.t.\ & \int_{\Omega} \Pr\{\mathbb{T} = 1|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_G(\mathbf{x}) d\mathbf{x} \leq s, \end{aligned} \tag{15}$$

where λ ($\lambda \geq 0$) is the Lagrange multiplier. Thus, under the screening setting with fairness constraint, the optimal choice for FAML is to admit candidates with characteristics $\mathbf{x}$ in a decreasing order of

$$f^0(\mathbf{x}) = \mathbb{E}_G(y|\mathbb{T} = 1, \mathbf{x}) + \lambda \cdot \Pr\{v = 1|\mathbb{T} = 1, \mathbf{x}\} \tag{16}$$

until reaching the capacity constraint.

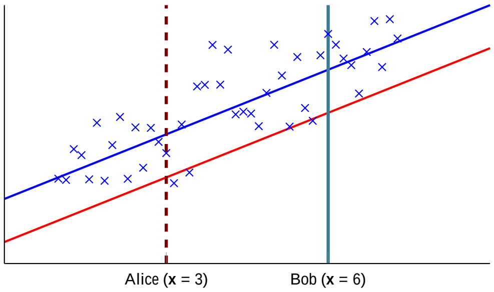
Juxtaposing Equation 16 with the optimal design for the selection case (i.e., Equation 13), the difference is, in essence, the same as the selection-screening difference without fairness constraint. That is, for screening, only candidates who can pass the manual interview matters for final selection quality and/or AIR. This is why Equation 16 includes $\mathbb{T} = 1$ as an additional condition compared with Equation 13. Note that, when a fairness constraint is present, the optimal outcome of manual interview can no longer be represented by a threshold cutoff on quality y (like in Equation 4). Instead, the optimal subset of candidates (who passed FAML screening) could feature different minimum quality for majority and minority candidates thanks to the fairness constraint. Thus, we now express the interview outcome as $\mathbb{T} = \mathbb{1}(y + \lambda_2 v \geq t_0)$, where $\mathbb{1}(\cdot)$ is again the indicator function, λ_2 captures the varying threshold between groups, and t_0 is the quality cutoff for the majority group (i.e., when $v = 0$). Taking this into Equation 16, we see that an FAML algorithm for screening would admit candidates in a decreasing order of

$$f^0(\mathbf{x}) = \mathbb{E}_G(y|y + \lambda_2 v \geq t_0, \mathbf{x}) + \lambda \cdot \Pr\{v = 1|y + \lambda_2 v \geq t_0, \mathbf{x}\} \tag{17}$$

until reaching the capacity constraint.

4.3 Comparison between Selection and Screening

We now examine how the design differences of FAML selection and screening algorithms could lead to different outcomes when both are used in the same setting – i.e., to retain s_1 fraction of candidates for manual interviews, which will eventually select s ($s \leq s_1$) fraction of candidates who must satisfy the fairness constraint of AIR $\geq r$. Again, both algorithms have access to the same training dataset and the same information (i.e., $\mathbf{x}$) about each candidate. Since the selection algorithm is unaware of the existence of manual


Alice (**x** = 3)
Bob (**x** = 6)

minorities are often less resourceful in preparing for such tests. To capture such between-group differences, we construct an applicant pool with equal fraction (i.e., 50%) of majority and minority candidates, and assign each group with the same quality distribution $N(5,1)$ but different $\mathbf{x}$-y relationship. Specifically, we calculate a real-valued $\mathbf{x}$ as a noisy proxy of y for each candidate,

$$\mathbf{x} = \begin{cases} y - 1 + \epsilon, & \text{if } v = 1 \text{ (i.e., minority)} \\ y + \epsilon, & \text{otherwise.} \end{cases} \tag{18}$$

where $\epsilon \sim N(0,1)$ is random noise. The resulting joint distribution Γ is depicted in Figure 2a. Note from the figure that minorities, on average, score lower on $\mathbf{x}$ than their majority counterparts of the same quality.

Before delving into the specifics of Alice and Bob, we first consider a well-recognized fairness issue associated with FAML selection algorithms which centers around the existence of within-group selection bias (Lipton et al. 2018, Zhang et al. 2023). Figure 2b depicts how the prediction target of FAML selection algorithms (i.e., $f(\mathbf{x})$ in Equation 13) varies with a candidate's characteristics $\mathbf{x}$ when the Lagrange multiplier $\lambda = 10$. The existence of within-group selection bias is evidenced by the non-monotonic nature of $f(\mathbf{x})$. On the one hand, note from Equation 18 that a larger $\mathbf{x}$ always implies a larger (expected value of) y for either majorities or minorities. On the other hand, the non-mononicity of $f(\mathbf{x})$ in Figure 2b suggests that an FAML selection algorithm, owing to its design of admitting candidates in a decreasing order of $f(\mathbf{x})$, could bypass a minority (or majority) candidate with a higher $\mathbf{x}$ (and hence a higher expected quality) to select another minority (or majority) with a lower $\mathbf{x}$ (i.e., a lower expected quality). This is the within-group selection bias recognized in existing work for FAML (Lipton et al. 2018, Zhang et al. 2023).

To explicate the reason behind this bias, and also to illustrate the difference between selection and screening, we consider how either FAML algorithm chooses between Alice with $\mathbf{x} = 3$ and Bob with $\mathbf{x} = 6$. Alice clearly has a lower expected quality $\mathbb{E}(y|\mathbf{x} = 3) = 3.68$ than Bob (6.32). Yet, as shown in Figure 2b, the FAML selection algorithm prefers Alice because her prediction target $f(\mathbf{x}) = 9.11$ is greater than Bob's (8.89). Figure 2c further illustrates why. The figure depicts the conditional probability density function of $z = y + \lambda \cdot v$ given $\mathbf{x}$ for Alice and Bob, respectively. Note that the prediction target for FAML selection algorithm is $f(\mathbf{x}) = \mathbb{E}(z|\mathbf{x})$, meaning that an FAML selection algorithm prefers candidates with a larger expected value of z. As can be seen from the figure, both Alice and Bob feature a bimodal distribution of z, with the left and right components corresponding to the case where the candidate is a majority and minority, respectively. Intuitively, as discussed earlier for Figure 1, the vertical height of the left component captures the *risk* associated with selecting a candidate, whereas the horizontal reach of the right component captures the potential *reward* from such a selection. From this perspective, it is clear that Bob is a high-risk high-reward choice because, even though both of its components have larger z than Alice, the risk of falling into the left component is considerably larger for Bob than for Alice. As a result, Alice has a larger expected value of z (9.11) than Bob (8.89), leading to her being preferred by the FAML selection algorithm. In other words, an FAML selection algorithm might skip a candidate with higher expected quality (i.e., Bob) simply because another candidate (i.e., Alice) looks more like a minority and is therefore a less risky choice (given the AIR constraint).

Figure 2d illustrates the case for FAML screening algorithm. As discussed in Section 4.2, for the screening algorithm, only candidates who can pass manual interview matters for either final selection quality or AIR. As such, the preference between Alice and Bob is now determined by the expected value of z for the non-shaded region only. Just like in the case without fairness constraint, this

5.1 Mathematical Analysis

A fairness constraint is only applicable when the distributions of predictors **x** or quality *y* differ between the majority and minority groups, because otherwise any selection algorithm $L(\mathbf{x})$ would produce the same selection rate for both groups. Thus, to analyze the outcome of FAML selection algorithm, we start with defining a measure of between-group difference according to the joint distribution G. Specifically, we are interested in between-group difference on $P(yj\mathbf{x})$, the conditional distribution of *y* given **x**, because FAML selection algorithm relies on $P(yj\mathbf{x})$ in their decision-making. To capture between-group difference on $P(yj\mathbf{x})$, we adopt a variation of Cohen's *d* (Cohen 2013), the standard statistic used in the US federal court system to establish a *prima facie* case of discrimination (Barnett 1982).

Definition 1 (Between-Group Difference). The between-group difference in G is defined as

$$\mathrm{d}_{\mathrm{G}} = \max_{\mathrm{Q}\quad\mathrm{W}} \frac{\mathbb{E}_{\mathrm{G}}(yj\mathbf{x}\ 2\ \mathrm{Q}; v = 0) - \mathbb{E}_{\mathrm{G}}(yj\mathbf{x}\ 2\ \mathrm{Q}; v = 1)}{}$$

shift our focus to cases with between-group difference (i.e., $d_G > 0$), and investigate the selection outcome when the ML algorithm is assigned with the selection task.

For the first step, we have the following theorem.

THEOREM 1. *For any joint distribution* G *with between-group difference* $d_G = 0$*, any selection rate* $s \in (0,1)$*, and any given fairness constraint AIR* $\geq r$ *(*$r \in [0,1]$*), there must exist a selection algorithm* $L(\mathbf{x})$ *that satisfies the selection rate s and fairness constraint AIR* $\geq r$ *while having selection quality* p_{SE} *matching the ideal value* p_{max}*. That is,*

$$p_{SE} = \max_{L:L\in\mathcal{L}} \int_W \mathbb{E}_G(y|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_G(\mathbf{x}) d\mathbf{x} = p_{max}, \tag{21}$$

where $\mathcal{L}$ *is the set of all possible selection algorithms that satisfy both capacity constraint s and AIR* $\geq r$*.*

As can be seen from the theorem, when G exhibits no between-group difference, then there would also be no within-group selection bias when assigning FAML with the selection task because the FAML selection algorithm can achieve the optimal selection quality p_{max}. For the second step, we have the following theorem.

THEOREM 2. *For any given probability density function of the predictor vector* $\mathbf{x}$*, any fairness constraint AIR* $\geq r$ *(*$r \in [0,1]$*), any selection rate* $s \in (0,1/2]$*, and any constant* $d > 0$*, there must exist a joint distribution* G *of predictor vector* $\mathbf{x}$*, group label v, and quality y, such that the between-group difference* $d_G \leq d$*, and*

$$\frac{p_{SE}}{p_{max}} \leq \frac{((2sr+1+r)^2 - 2sr(1+r)d^2) \cdot r \cdot (1+r-2s) + (2sr+1+r)^3}{(1+r-2s)r(1+r)^2d^2 + (1+r)^2(2sr+1+r)^2}. \tag{22}$$

When $s \to 0$*, the limit of this ratio satisfies*

$$\lim_{s\to 0} \frac{p_{SE}}{p_{max}} \leq \frac{r+1}{rd^2+r+1}. \tag{23}$$

Consistent with our earlier conceptual development, Theorem 2 shows that, when between-group difference is present, assigning ML with the selection task necessitates a deviation from quality-based selection and results in a substantial loss of selection quality. For example, even when the between-group bias is quite small, e.g., $d_G \leq 0.5$, to achieve AIR ≥ 0.8, we have $p_{SE}/p_{max} \leq (0.8+1)/(0.8 \cdot 0.25 + 0.8 + 1) = 0.9$ when $s \to 0$, suggesting a loss of at least 10% on selection quality. When the between-group difference is larger, e.g., $d_G = 1$, there is $p_{SE}/p_{max} \leq 0.69$ when $s \to 0$, indicating a loss of over 30% for selection quality. Further, the theorem also shows that the upper bound on p_{SE}/p_{max} decreases with a larger[7] r, indicating that the problem with the selection task becomes more severe when the fairness constraint is more stringent. These results confirm our earlier observations that, with the presence of between-group difference, assigning ML with the selection task could lead to a departure from quality-based selection, resulting in within-group selection bias and, consequently, a substantial decrease in final selection quality. This demonstrates the importance of building manual examination (e.g., interviews) into selection processes in practice.

[7] Note that the partial derivative of $\lim_{s\to 0} p_{SE}/p_{max}$ with respect to r is $-d^2/(rd^2+r+1)^2 \leq 0$.

5.2 Simulation Study

In this subsection, we present a simulation study that compares the outcomes of 1) directly using an FAML algorithm for selection; and 2) using an FAML algorithm for screening followed by manual interviews. We describe the dataset, the design of the simulation study, and the results, respectively.

5.2.1 Dataset

While our findings apply to a wide variety of selection settings, from college admissions to loan applications, among them personnel selection is a setting that has received the most empirical attention in the literature (SIOP 2018). We thus designed our simulation study by following the prevailing practice in personnel selection (Finch et al. 2009), which is to construct a dataset according to the empirical evidence reported in meta-analysis (Bobko et al. 1999) pertaining to the 1) the correlation between predictor variables and the quality indicator, 2) the inter-correlation among predictor variables, and 3) the between-group difference on each predictor.

Table 1 Standardized Mean Group Differences and Correlation Matrix

Variable	1	2	3	4	5	6	*d*
1. Biodata	–						0.33
2. Cognitive ability	.19	–					1.00
3. Conscientiousness	.51	.00	–				0.09
4. Integrity	.25	.00	.39	–			0.00
5. Structured interview	.16	.24	.12	.00	–		0.23
6. Job performance (*y*)	.28	.30	.18	.25	.30	–	0.45

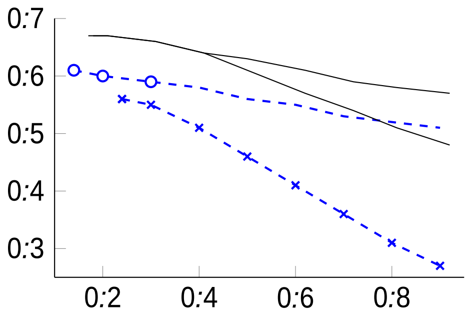
Note. Variables 1-4 = **x**, predictors available to ML. Variable 5 = predictor administered manually post-screening (if applicable). Variable 6 = quality indicator *y*. *d* = standardized mean group difference between Black and White applicants.

To this end, we followed the exact same procedure as e 6(most)-3mt(as)-2922(1s)-2f ws an excal exadence

5.2.2 Design

Table 2 Mean Quality of Selected Candidates When AIR = .3, .6, .9

	$s = :10$											$s = :20$										
p_1,	Selection (s_1=s = 1)			Screening (s_1=s = 2)				Screening (s_1=s = 3)				Selection (s_1=s = 1)			Screening (s_1=s = 2)				Screening (s_1=s = 3)			
AIR	ID	ML	d_1	ID	ML	d_1	d_2	ID	ML	d_1	d_2	ID	ML	d_1	ID	ML	d_1	d_2	ID	ML	d_1	d_2
.2,.3	0.71	0.68	.04	0.78	0.77	.00	.01	0.79	0.79	.00	.00	0.57	0.55	.03	0.63	0.63	.00	.00	0.63	0.63	.00	.00
	(.02)	(.02)	(.01)	(.02)	(.02)	(.00)	(.00)	(.02)	(.02)	(.00)	(.00)	(.08)	(.08)	(.00)	(.08)	(.08)	(.00)	(.00)	(.09)	(.09)	(.00)	(.00)
.2,.6	0.69	0.59	.15	0.76	0.73	.00	.05	0.77	0.77	.00	.01	0.55	0.47	.15	0.61	0.60	.00	.01	0.61	0.61	.00	.00
	(.02)	(.02)	(.01)	(.03)	(.03)	(.00)	(.01)	(.02)	(.03)	(.00)	(.01)	(.08)	(.06)	(.01)	(.08)	(.08)	(.00)	(.01)	(.08)	(.08)	(.00)	(.01)
.2,.9	0.67	0.49	.28	0.73	0.68	.00	.08	0.75	0.72	.00	.03	0.53	0.38	.29	0.59	0.57	.00	.03	0.59	0.59	.00	.00
	(.02)	(.02)	(.01)	(.03)	(.03)	(.00)	(.01)	(.03)	(.03)	(.00)	(.01)	(.08)	(.06)	(.02)	(.08)	(.08)	(.00)	(.01)	(.08)	(.08)	(.00)	(.01)
.4,.3	0.67	0.62	.07	0.74	0.74	.00	.00	0.74	0.74	.00	.00	0.51	0.48	.06	0.56	0.56	.00	.00	0.56	0.56	.00	.00
	(.02)	(.02)	(.01)	(.02)	(.02)	(.00)	(.00)	(.02)	(.02)	(.00)	(.00)	(.08)	(.08)	(.00)	(.08)	(.08)	(.00)	(.00)	(.09)	(.09)	(.00)	(.00)
.4,.6	0.64	0.49	.23	0.71	0.67	.00	.05	0.71	0.71	.00	.01	0.49	0.39	.21	0.54	0.53	.00	.01	0.54	0.54	.00	.00
	(.02)	(.02)	(.02)	(.02)	(.02)	(.00)	(.01)	(.02)	(.02)	(.00)	(.01)	(.08)	(.06)	(.02)	(.08)	(.07)	(.00)	(.01)	(.08)	(.08)	(.00)	(.01)
.4,.9	0.61	0.36	.41	0.67	0.60	.00	.09	0.67	0.65	.00	.04	0.45	0.26	.42	0.50	0.48	.00	.05	0.51	0.50	.00	.01
	(.02)	(.02)	(.03)	(.02)	(.03)	(.02)	(.02)	(.02)	(.03)	(.00)	(.02)	(.08)	(.06)	(.03)	(.08)	(.07)	(.01)	(.02)	(.08)	(.07)	(.00)	(.02)
.6,.3	0.59	0.55	.08	0.66	0.66	.00	.00	0.66	0.66	.00	.00	0.43	0.39	.08	0.48	0.48	.00	.00	0.48	0.48	.00	.00
	(.02)	(.02)	(.01)	(.02)	(.02)	(.00)	(.00)	(.02)	(.02)	(.00)	(.00)	(.08)	(.08)	(.01)	(.08)	(.08)	(.00)	(.00)	(.09)	(.09)	(.00)	(.00)
.6,.6	0.55	0.41	.26	0.61	0.57	.00	.06	0.62	0.61	.00	.01	0.39	0.30	.22	0.44	0.44	.00	.00	0.45	0.45	.00	.00
	(.02)	(.02)	(.02)	(.02)	(.02)	(.00)	(.01)	(.02)	(.02)	(.00)	(.01)	(.08)	(.06)	(.03)	(.08)	(.07)	(.00)	(.02)	(.08)	(.08)	(.00)	(.01)
.6,.9	0.51	0.27	.46	0.57	0.48	.07	.16	0.57	0.52	.00	.08	0.35	0.17	.50	0.40	0.37	.00	.09	0.41	0.40	.00	.02
	(.02)	(.02)	(.03)	(.02)	(.03)	(.03)	(.03)	(.02)	(.03)	(.00)	(.02)	(.08)	(.05)	(.05)	(.08)	(.06)	(.03)	(.03)	(.08)	(.07)	(.00)	(.03)
avg	0.63	0.50	.22	0.69	0.66	.01	.06	0.70	0.69	.00	.02	0.47	0.38	.22	0.53	0.52	.00	.02	0.53	0.53	.00	.00


0:7
0:6
0:5
0:4
0:3
0:2
0:4
0:6
0:8

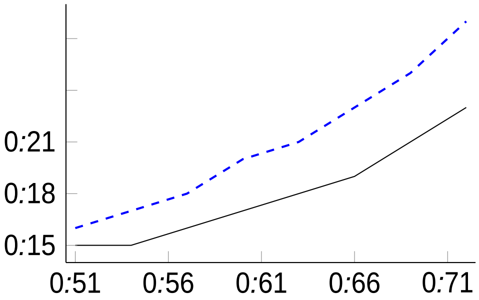
5.3.2 Design of ML Algorithms

To ensure a fair comparison, for each dataset, we used the exact same ML algorithm for selection and screening, with the only exception being their respective prediction targets as defined in Equations 13 and 16, respectively. For the simulation dataset, since the variables were generated as a mixture of multivariate Gaussian distributions, the natural choice for ML algorithm is the iterative Expectation-Maximization (EM) algorithm for learning a Gaussian mixture model (McLachlan et al. 2019). For the real-world dataset, the high dimensionality of $\mathbf{x}$ (i.e., 120 variables) could easily lead to curse-of-dimensionality problems for many ML algorithms (Bengio and Bengio 2000), e.g., support vector machines, Gaussian processes, etc. To address the challenge, we used a multilayer perceptron (MLP; Goodfellow et al. 2016) – i.e., a feed-forward, fully connected neural network – which is known to excel at handling high-dimensional data (Poggio et al. 2017). It is important to note, however, that our choice of using MLP in this context is for demonstration purposes only, and should not be interpreted as a suggestion of its superiority over other alternative algorithms (e.g., regularized regression). Specifically, we trained a simple MLP with three layers, a hidden layer size of 10, and the Rectified Linear Unit (ReLU) activation function following each layer except the last (Goodfellow et al. 2016). Given the vast scale difference of different predictors, we followed the common standardization procedure (i.e., using z-score) for each variable before feeding data into the MLP. The training of MLP was done using the limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) algorithm (Nocedal and Wright 2006) to minimize the mean squared error of predictions.

5.3.3 Results

For both datasets, we tested the selection and screening algorithms with a final selection rate of $s = 0{:}1$ and a fairness constraint of AIR ≥ 1. Both algorithms were used to retain s_1 ($s_1 > s$) fraction of candidates, who are then further selected through manual interviews that are implemented in the exact same way for both algorithms. Specifically, to ensure that any degradation of selection quality can be attributed to the ML algorithms rather than the manual interviews, we set the interviews to generate the optimal outcome for both algorithms – i.e., to select the subset of retained candidates with the highest expected quality, subject to capacity (i.e., s) and fairness (i.e., AIR ≥ 1) constraints.

With this setup, there is clearly a tradeoff between s_1 and the final selection quality $\bar{y}$ (i.e., the average quality of all s selected candidates) for both algorithms, because either algorithm could achieve the same, best possible, selection quality when $s_1 = 1$. We denote such best possible quality as $\bar{y}_{\max}$. To assess the tradeoff achieved by the two algorithms, we varied the retention rate s_1 from 0.15 to 0.30 (with a step of 0.01), and then compared the minimum retention rate s_1 required by either algorithm to reach a certain fraction (e.g., 80%) of the best possible quality $\bar{y}_{\max}$. Clearly, this comparison would directly reveal the saving of interview cost should we replace one algorithm with the other.


0:21
0:18
0:15
0:51
0:56
0:61
0:66
0:71

research may examine how such data- and algorithm-quality issues could affect the outcomes of FAML algorithms in selection and screening settings.

We also offer the caveat that the current work was situated in the legal context in the US. We did not consider the egalitarian ideals of fairness, despite its popularity in FAML research as the basis of fairness definitions (Mitchell et al. 2018). We also did not consider the perception of fairness, such as whether the use of algorithms for selection could undermine individual's beliefs about procedural justice (Newman et al. 2020). While the selection-screening distinction studied in the paper is a fundamental issue that transcends national boundaries, the specific legal environment could differ drastically from one country to another (Sánchez-Monedero et al. 2020). Thus, our results may be less applicable to nations where anti-discrimination laws do not stipulate limits on disparate impact, hence rendering the enforcement of fairness constraints less relevant (Mahlmann 2015, Murphy 2018).

Finally, we focused on AIR as the fairness measure in this paper because of its widespread use in the US legal system. In the FAML literature, many other measures have been studied (Mitchell et al. 2018). They range from statistical parity (between groups) on selection rates (Zemel et al. 2013, Agarwal et al. 2018) to statistical parity on predictive accuracy (Feldman et al. 2015, Donini et al. 2018), from a constraint on mapping similar predictors to similar outcomes (e.g., Lipschitz constraint; Dwork et al. 2012; no preferential treatment; Joseph et al. 2016) to an assurance that no protected group under one selection system would overwhelmingly prefer another system (i.e., "envy-freeness"; Zafar et al. 2019, Ustun et al. 2019), from a measure specified through causal or counterfactual inference (Datta et al. 2017, Kilbertus et al. 2017, Kusner et al. 2017, Nabi and Shpitser 2018, Zhang and Bareinboim 2018) to a combination of multiple constraints (Hardt et al. 2016). These constraints are so diverse that, as noted repeatedly in the FAML literature (Kleinberg et al. 2017, Chouldechova 2017, Pleiss et al. 2017), many of them are inherently conflicted even without considering selection quality. Future research may examine how the use of other fairness constraints may affect the difference between selection and screening tasks for FAML algorithms.

Acknowledgement

This work was supported in part by the National Science Foundation and by gifts from Amazon Science and Meta. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudik, John Langford, Hanna Wallach. 2018. A reductions approach to fair classification. *Proceedings of Machine Learning Research*, 80 60-69.

Aguinis, Herman, Steven A Culpepper, Charles A Pierce. 2010. Revival of test bias research in preemployment testing. *Journal of Applied Psychology*, 95 (4), 648-680.

Aksin, Zeynep, Mor Armony, Vijay Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16 (6), 665-688.

Arlotto, Alessandro, Stephen E Chick, Noah Gans. 2014. Optimal hiring and retention policies for heterogeneous workers who learn. *Management Science*, 60 (1), 110-129.

Barnett, Arnold. 1982. An underestimated threat to multiple regression analyses used in job discrimination cases. *Industrial Relations Law Journal*, 5 156.

Bengio, Samy, Yoshua Bengio. 2000. Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks*, 11 (3), 550-557.

Bent, Jason R. 2020. Is algorithmic affirmative action legal? *Georgetown Law Journal*, 108 (4), 803-853.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, Advance online publication. `https://doi.org/10.1177/0049124118782533`.

Berry, Christopher M, Malissa A Clark, Tara K McClure. 2011. Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology*, 96 (5), 881.

Bobko, Philip, Philip L Roth, Denise Potosky. 1999. Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel psychology*, 52 (3), 561-589.

Boudreau, John, Wallace Hopp, John O McClain, L Joseph Thomas. 2003. On the interface between operations and human resources management. *Manufacturing & Service Operations Management*, 5 (3), 179-202.

Burke, Lilah. 2020. The death and life of an admissions algorithm. Inside Higher Ed. `https://insidehighered.com/admissions/article/2020/12/14/u-texas-will-stop-using-controversial-algorithm-evaluate-phd`.

Chouldechova, Alexandra. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5 (2), 153-163.

Cohen, Jacob. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

on Computer and Communications Security. 1193-1210.

De-Arteaga, Maria, Stefan Feuerriegel, Maytal Saar-Tsechansky. 2022. Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management*, 31 (10), 3749-3770.

De Corte, Wilfried, Paul R Sackett, Filip Lievens. 2011. Designing pareto-optimal selection systems: Formalizing the decisions required for selection system development. *Journal of Applied Psychology*, 96 (5), 907-926.

Donini, Michele, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. *Advances in Neural Information Processing Systems*, 31 2796-2806.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard Zemel. 2012. Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214-226.

Dwork, Cynthia, Nicole Immorlica, Adam Tauman Kalai, Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. *Proceedings of Machine Learning Research*, 81 119-133.

Eriksson, Kimmo, Jonas Sjöstrand, Pontus Strimling. 2007. Optimal expected rank in a two-sided secretary problem. *Operations Research*, 55 (5), 921-931.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 259-268.

Finch, David M, Bryan D Edwards, J Craig Wallace. 2009. Multistage selection strategies: Simulating the effects on adverse impact and expected performance for various predictor combinations. *Journal of Applied Psychology*, 94 (2), 318-340.

Fu, Runshan, Manmohan Aseri, Param Vir Singh, Kannan Srinivasan. 2022. "un" fair machine learning algorithms. *Management Science*, 68 (6), 4173-4195.

Fu, Runshan, Yan Huang, Param Vir Singh. 2021. Crowds, lending, machine, and bias. *Information Systems Research*, 32 (1), 72-92.

Fuller, Mercedes, Paul Swiontkowski. 2020. The AI revolution is coming. Accenture-Microsoft Report, `https://www.accenture.com/us-en/insights/software-platforms/ai-revolution-coming`. Accessed: 2020-10-07.

Gikay, Asress Adimi. 2020. The american way-until machine learning algorithm beats the law? *Case W. Res. JL Tech. & Internet*, 12 ii.

Gonzalez, Manuel F, John F Capman, Frederick L Oswald, Evan R Theys, David L Tomczak. 2019. "where's the io?" artificial intelligence and machine learning in talent management systems. *Personnel Assessment and Decisions*, 5 (3), 33-44.

Goodfellow, Ian, Yoshua Bengio, Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gottfredson, Linda S. 1994. The science and politics of race-norming. *American Psychologist*, 49 (11), 955.

Hardt, Moritz, Eric Price, Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29 3315-3323.

Hardy, Godfrey Harold, John Edensor Littlewood, George Pólya. 1952. *Inequalities*. Cambridge university press.

Hough, Leaetta M, Frederick L Oswald. 2000. Personnel selection: Looking toward the future–remembering the past. *Annual Review of Psychology*, 51 (1), 631-664.

Hunter, John E, Ronda F Hunter. 1984. Validity and utility of alternative predictors of job performance. *Psychological bulletin*, 96 (1), 72-98.

Joseph, Matthew, Michael Kearns, Jamie H Morgenstern, Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems*, 29 325-333.

Kallus, Nathan, Xiaojie Mao, Angela Zhou. 2022. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68 (3), 1959-1981.

Kendall, Alex, Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30.

Lipton, Zachary, Julian McAuley, Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity? *Advances in Neural Information Processing Systems*, 31 8125-8135.

Louizos, Christos, Kevin Swersky, Yujia Li, Max Welling, Richard S Zemel. 2016. The variational fair autoencoder. *Proceedings of the International Conference on Learning Representations*.

Mahlmann, M. 2015. Country report, non-discrimination, germany. *European network of legal experts in gender equality and non-discrimination, Directorate-General for Justice and Consumers, Publications Office of the European Union, Luxembourg*, .

Martinez, Emmanuel, Lauren Kirchner. 2021. The secret bias hidden in mortgage-approval algorithms. The Markup. `https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms`. Accessed: 2022-10-07.

McLachlan, Geoffrey J, Sharon X Lee, Suren I Rathnayake. 2019. Finite mixture models. *Annual review of statistics and its application*, 6 355-378.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6), 1-35.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, .

Mithas, Sunil, Zhi-Long Chen, Terence JV Saldanha, Alysson De Oliveira Silveira. 2022. How will artificial intelligence and industry 4.0 emerging technologies transform operations management? *Production and Operations Management*, in press.

Morgeson, Frederick P, Michael A Campion, Robert L Dipboye, John R Hollenbeck, Kevin Murphy, Neal Schmitt. 2007. Reconsidering the use of personality tests in personnel selection contexts. *Personnel psychology*, 60 (3), 683-729.

Murphy, Kevin R. 2018. The legal context of the management of human resources. *Annual Review of Organizational Psychology and Organizational Behavior*, 5 157-182.

Nabi, Razieh, Ilya Shpitser. 2018. Fair inference on outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*. 1931-1940.

chap. 5. Routledge, New York and London, 92-112.

Pedreshi, Dino, Salvatore Ruggieri, Franco Turini. 2008. Discrimination-aware data mining. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 560-568.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in Neural Information Processing Systems*, 30 5680-5689.

Poggio, Tomaso, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, Qianli Liao. 2017. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14 (5), 503-519.

Primus, Richard. 2010. The future of disparate impact. *Michigan Law Review*, 108 1341-1387.

Purkiss, Sharon L Segrest, Pamela L Perrewé, Treena L Gillespie, Bronston T Mayes, Gerald R Ferris. 2006. Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes*, 101 (2), 152-167.

Rambachan, Ashesh, Jon Kleinberg, Sendhil Mullainathan, Jens Ludwig. 2020. An economic approach to regulating algorithms. Tech. rep., National Bureau of Economic Research.

Rasmussen, C, C Williams. 2006. *Gaussian processes for machine learning*. MIT Press.

Roth, Philip L, Allen I Huffcutt, Philip Bobko. 2003. Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology*, 88 (4), 694.

Sackett, Paul R, Steffanie L Wilk. 1994. Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, 49 (11), 929-954.

Sánchez-Monedero, Javier, Lina Dencik, Lilian Edwards. 2020. What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 458-468.

SIOP. 2018. *Principles for the validation and use of personnel selection procedures*. 5th ed. SIOP.

Song, Q, Serena Wee, Daniel A Newman. 2017. Diversity shrinkage: Cross-validating pareto-optimal weights to enhance diversity via hiring practices. *Journal of Applied Psychology*, 102 (12), 1636-1657.

Ustun, Berk, Yang Liu, David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. *Proceedings of Machine Learning Research*, 97 6373-6382.

Wang, Hao, Hsiang Hsu, Mario Diaz, Flavio P Calmon. 2021. To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, 67 (10), 6733-6757.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P Gummadi. 2019. Fairness constraints:

Supplemental Materials

EC.1 Proof of Theorem 1

THEOREM 1. *For any joint distribution* G *with between-group difference* $d_G = 0$*, any selection rate* $s\ 2\ (0; 1)$*, and any given fairness constraint* $AIR \geq r\ (r\ 2\ [0; 1])$*, there must exist a selection algorithm* L *that satisfies the selection rate* s *and fairness constraint* $AIR \geq r$ *while having selection quality* p_{SE} *matching the ideal value* p_{max}*. That is,*

$$p_{SE} = \max_{L:L2\mathcal{L}} \underset{W}{\overset{Z}{}} \mathbb{E}_G(yj\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_G(\mathbf{x})d\mathbf{x} = p_{max}; \qquad \text{(EC.1)}$$

where $L(\mathbf{x})$ *is the probability for* L *to select a candidate with predictor vector* $\mathbf{x}$*, and* $\mathcal{L}$ *is the set of all possible selection algorithms that satisfy both capacity constraint* s *and* $AIR \geq r$*.*

Proof. The proof is structured as follows. We start by considering the optimal algorithm T that makes use of the group label v $(v\ 2\ f0; 1g)$ of each candidate to achieve the ideal selection quality p_{max}. Note that the existence of T is evident from the definition of p_{max} in Equation 20. Since T has access to the group label, we use $T(\mathbf{x}; v)$ to denote[9] the probability for T to select a candidate with predictor vector $\mathbf{x}$ and group label v. Then, we construct a selection algorithm L, which does *not* use v in its input, based on T and prove that L matches T in terms of selection rate, selection quality, and AIR.

Our construction of L is quite simple. For any given candidate with predictor vector $\mathbf{x}\ 2$ W, we set L to select the candidate with probability

$$L(\mathbf{x}) = T(\mathbf{x}; 1); \qquad \text{(EC.2)}$$

where $T(\mathbf{x}; 1)$ is the selection probability, according to T, for a candidate with predictor vector $\mathbf{x}$ and group label $v = 1$ (i.e., minority).

We prove by contradiction that L matches T in terms of selection rate, selection quality, and AIR. Suppose they do not match. Then there must exist $\mathbf{x}\ 2$ W such that $T(\mathbf{x}; 0)\ 6= T(\mathbf{x}; 1)$, because otherwise L and T would be equivalent for all candidates (i.e., $8\mathbf{x}\ 2$ W, $L(\mathbf{x}) = T(\mathbf{x}; 0) = T(\mathbf{x}; 1)$).

Given the existence of $\mathbf{x}$ with $T(\mathbf{x}; 0)\ 6= T(\mathbf{x}; 1)$, one of following two possibilities must be true: 1) there exists at least one predictor vector $\mathbf{x}_0\ 2$ W such that $T(\mathbf{x}_0; 0) > T(\mathbf{x}_0; 1)$; and 2) there is $T(\mathbf{x}; 0) \leq T(\mathbf{x}; 1)$ for all $\mathbf{x}\ 2$ W, with the inequality (i.e., $<$) holding for at least some $\mathbf{x}$. We consider the two cases respectively in the rest of 10.9589 10.9589 Tf 5.479 0 Td [(j979589 yS [(x)]TJ/F124 10.9589 Tf 5.479 0 Td [(;)]TJ/F88 10.9589 Tf 4.558

Case 1: For the first case, we prove contradiction by first constructing an alternative algorithm T' that also makes use of group label v but is different from T, and then proving that T cannot be optimal (i.e., contradiction) because T' dominates it in terms of the tradeoff between selection quality and AIR.

Specifically, recall that, in this first case, there exists $\mathbf{x}_0 \in W$ such that $T(\mathbf{x}_0, 0) > T(\mathbf{x}_0, 1)$. We set $T'(\mathbf{x}, v) = T(\mathbf{x}, v)$ for all $\mathbf{x} \in W \setminus \{\mathbf{x}_0\}$ and $v \in \{0, 1\}$. For $\mathbf{x} = \mathbf{x}_0$, we set

$$T'(\mathbf{x}_0, 1) = T(\mathbf{x}_0, 0) \cdot \Pr\{v = 0|\mathbf{x}_0\} + T(\mathbf{x}_0, 1) \cdot \Pr\{v = 1|\mathbf{x}_0\} \tag{EC.3}$$

$$> T(\mathbf{x}_0, 1), \text{ and} \tag{EC.4}$$

$$T'(\mathbf{x}_0, 0) = T(\mathbf{x}_0, 0) \cdot \Pr\{v = 0|\mathbf{x}_0\} + T(\mathbf{x}_0, 1) \cdot \Pr\{v = 1|\mathbf{x}_0\} \tag{EC.5}$$

$$< T(\mathbf{x}_0, 0). \tag{EC.6}$$

The inequalities in (EC.4) and (EC.6) hold because $T(\mathbf{x}_0, 0) > T(\mathbf{x}_0, 1)$. A key observation from the construction of T' is that

$T(\mathbf{x}, 0) \cdot \Pr\{v = 0= 0= 0==0=$

meaning that T yields a selection outcome with $\mathrm{AIR} > 1$, contradicting our assumption of $\mathrm{AIR} \in [0, 1]$.

Step 2: As Step 1 proves the existence of D, the objective of Step 2 is to prove that there must exist a pair of predictor vectors $\mathbf{x}_0$ and $\mathbf{x}_0'$, one inside and the other outside D, with different expected quality. In other words, there must exist $\mathbf{x}_0 \in D$ and $\mathbf{x}_0' \in W \backslash D$, such that

$$\mathbb{E}_G(y|\mathbf{x}_0) \neq \mathbb{E}_G(y|\mathbf{x}_0'). \tag{EC.12}$$

The reason here is straightforward. If no such pair exists, then all $\mathbf{x}$ ($\mathbf{x} \in W$) must share the same expected quality according to G. Given $d_G = 0$, any selection outcome would then yield the exact same selection quality, directly proving the theorem[10].

Step 3: As Step 2 proves the existence of $\mathbf{x}_0 \in D$ and $\mathbf{x}_0' \in W \backslash D$ with $\mathbb{E}_G(y|\mathbf{x}_0) \neq \mathbb{E}_G(y|\mathbf{x}_0')$, we are now ready to complete the proof for Case 2. Consider the between-group difference for where the inequality in (EC.20) is due to (EC.8) and $\mathbf{x}_0 \in \mathrm{D}$, and the inequality in (EC.21) is due to $\mathbf{x}_0' \notin \mathrm{D}$. Taking the result here (i.e., $\Pr\{\mathbf{x} = \mathbf{x}_0 | \mathbf{x} \in \{\mathbf{x}_0, \mathbf{x}_0'\}, v = 0\} - \Pr\{\mathbf{x} = \mathbf{x}_0 | \mathbf{x} \in \{\mathbf{x}_0, \mathbf{x}_0'\}, v = 1\} > 0$) and (EC.12) into (EC.16), we have

$$\mathbb{E}_{\mathrm{G}}(y | \mathbf{x} \in \{\mathbf{x}_0, \mathbf{x}_0'\}, v = 0) - \mathbb{E}_{\mathrm{G}}(y | \mathbf{x} \in \{\mathbf{x}_0, \mathbf{x}_0'\}, v = 1) \neq 0, \tag{EC.24}$$

which contradicts[11] the assumption that $\mathrm{d}_{\mathrm{G}} = 0$. This completes the proof for both cases.

[11] To see this, simply take $\mathrm{Q} = \{\mathbf{x}_0, \mathbf{x}_0'\}$ in Definition 1.

EC.2 Proof of Theorem 2

THEOREM 2 *For any given probability density function of the predictor vector* $\mathbf{x}$, *any fairness constraint AIR* $\geq r$ *(*$r \in [0,1]$*), any selection rate* $s \in (0,1/2]$, *and any constant* $d > 0$, *there must exist a joint distribution* G *of predictor vector* $\mathbf{x}$, *group label* v, *and quality* y, *such that the between-group difference* $\mathrm{d}_{\mathrm{G}} \leq d$, *and*

$$\frac{\mathrm{p}_{\mathrm{SE}}}{\mathrm{p}_{\max}} \leq \frac{((2sr+1+r)^2 - 2sr(1+r)d^2) \cdot r \cdot (1+r-2s) + (2sr+1+r)^3}{(1+r-2s)r(1+r)^2 d^2 + (1+r)^2(2sr+1+r)^2}. \tag{EC.25}$$

When $s \to 0$, *the limit of this ratio satisfies*

$$\lim_{s \to 0} \frac{\mathrm{p}_{\mathrm{SE}}}{\mathrm{p}_{\max}} \leq \quad r +$$

Upper bound on d_G*:* Recall from Definition 1 that, to prove $\mathrm{d}_G \le d$, we need to prove

$$\frac{\mathbb{E}_G(y|\mathbf{x}\in Q; v=0) - \mathbb{E}_G(y|\mathbf{x}\in Q; v=1)}{\mathrm{SD}_G(y|\mathbf{x}\in Q)} \le d \tag{EC.42}$$

for all possible $Q \subseteq W$. Note from our construction of μ_{ij} that for all $\mathbf{x}\in W$, there is $\mathbb{E}_G(y|\mathbf{x}; v=0) = \mu$. This reduces (EC.42) to

$$\frac{\mu - \mathbb{E}_G(y|\mathbf{x}\in Q; v=1)}{\mathrm{SD}_G(y|\mathbf{x}\in Q)} \le d. \tag{EC.43}$$

Further, note that for all $x \notin W_1$, there is $\mathbb{E}_G(y|\mathbf{x}; v=1) = \mu$. Thus, we only need to consider $Q \subseteq W_1$ when proving (EC.42). Since both the conditional distribution of v given $\mathbf{x}$ and the conditional distribution of y given $\mathbf{x}; v$ stay constant for all $\mathbf{x}\in W_1$, we only need to prove

$$\frac{\mu - \mathbb{E}_G(y|\mathbf{x}\in W_1; v=1)}{\mathrm{SD}_G(y|\mathbf{x}\in W_1)} \le d. \tag{EC.44}$$

As $\mathbb{E}_G(y|\mathbf{x}\in W_1; v=1) = \mu_{11} = 1$, our objective now is to prove

$$\frac{1-\mu}{\mathrm{SD}_G(y|\mathbf{x}\in W_1)} \le d. \tag{EC.45}$$

To calculate $\mathrm{SD}_G(y|\mathbf{x}\in W_1)$, note that the conditional distribution of y given $\mathbf{x}\in W_1$ follows Bernoulli distribution with mean

$$\begin{aligned}
\mu_1 &= \Pr\{v=0|\mathbf{x}\in W_1\}\cdot\mu_{10} + \Pr\{v=1|\mathbf{x}\in W_1\}\cdot\mu_{11} && \text{(EC.46)}\\
&= \frac{1+r}{2sr+1+r}\cdot\mu + \frac{2sr}{2sr+1+r}\cdot 1 && \text{(EC.47)}\\
&= \frac{2sr+\mu(1+r)}{2sr+1+r}. && \text{(EC.48)}
\end{aligned}$$

By taking the definition of μ in (EC.41) into (EC.48), we have

$$\begin{aligned}
\mu_1 &= \max\left(\frac{2sr}{2sr+1+r}, \frac{2sr + \frac{(2sr+1+r)^2 - 2sr(1+r)d^2}{(1+r)^2d^2+(2sr+1+r)^2}\cdot(1+r)}{2sr+1+r}\right) && \text{(EC.49)}\\
&= \max\left(\frac{2sr}{2sr+1+r}, \frac{2sr + \frac{(1+r)(2sr+1+r)^2 - 2sr(1+r)^2d^2}{(1+r)^2d^2+(2sr+1+r)^2}}{2sr+1+r}\right) && \text{(EC.50)}\\
&= \max\left(\frac{2sr}{2sr+1+r}, \frac{\frac{(2sr+1+r)^3}{(1+r)^2d^2+(2sr+1+r)^2}}{2sr+1+r}\right) && \text{(EC.51)}\\
&= \max\left(\frac{2sr}{2sr+1+r}, \frac{(2sr+1+r)^2}{(1+r)^2d^2+(2sr+1+r)^2}\right). && \text{(EC.52)}
\end{aligned}$$

We are now ready to prove (EC.45). Specifically, we calculate its left-hand side as

$$
\begin{aligned}
\frac{1-\mu}{\mathrm{SD}_{\mathrm{G}}(y \mid \mathbf{x} \in \mathrm{W}_1)} &= \frac{1-\mu}{\sqrt{\mu_1 \cdot (1-\mu_1)}} && \text{(EC.53)}\\
&= \frac{1-\mu}{\sqrt{\mu_1 \cdot \left(1 - \frac{2sr+\mu(1+r)}{2sr+1+r}\right)}} && \text{(EC.54)}\\
&= \frac{1-\mu}{\sqrt{\mu_1 \cdot \frac{(1-\mu)(1+r)}{2sr+1+r}}} && \text{(EC.55)}\\
&= \sqrt{\frac{1-\mu}{\mu_1 \cdot \frac{1+r}{2sr+1+r}}} && \text{(EC.56)}\\
&\leq \sqrt{\frac{1-\mu}{\frac{(2sr+1+r)^2}{(1+r)^2d^2+(2sr+1+r)^2} \cdot \frac{1+r}{2sr+1+r}}}; && \text{(EC.57)}
\end{aligned}
$$

where (EC.54) is derived by replacing the second appearance of μ_1 with the expression in (EC.48); (EC.56) is derived by dividing both the numerator and the denominator by $\sqrt{1-\mu}$; and the inequality in (EC.57) is derived by replacing μ_1 with the second term in (EC.52).

Recall from (EC.41) that μ takes the maximum value of 0 or $\frac{(2sr+1+r)^2-2sr(1+r)d^2}{(1+r)^2d^2+(2sr+1+r)^2}$. If $\mu = 0$, there must be

This completes the proof of (EC.45) and, in turn, $\mathrm{d}_{\mathrm{G}} \leq d$.

Upper bound on p_{SE}=$\mathrm{p}_{\mathrm{max}}$*:* First, consider the ideal algorithm with access to v, which has an expected selection quality of $\mathrm{p}_{\mathrm{max}}$. Note that, with our construction of G, the ideal algorithm could simply choose all minority candidates with $\mathbf{x}$ *2* W_1 and $v = 1$. Since

$$\mathrm{Pr} f\mathbf{x}\ 2\ \mathrm{W}_1; v = 1g = \mathrm{Pr} fv = 1 j\mathbf{x}\ 2\ \mathrm{W}_1 g \cdot \mathrm{Pr} f\mathbf{x}\ 2\ \mathrm{W}_1 g \qquad \text{(EC.66)}$$

$$= \frac{2sr}{2sr + 1 + r} \cdot \quad \frac{1}{2} + \frac{sr}{1 + r} \qquad \text{(EC.67)}$$

$$= \frac{2sr}{2sr + 1 + r} \cdot \frac{2sr + 1 + r}{2(1 + r)} \qquad \text{(EC.68)}$$

$$= \frac{sr}{1 + r}; \qquad \text{(EC.69)}$$

selecting all minority candidates with $\mathbf{x}$ *2* W_1 exactly reaches AIR $= r$. We therefore have

$$\mathrm{p}_{\mathrm{max}} = \mu_{11} \cdot \frac{r}{1 + r} + \mu \cdot \frac{1}{1 + r} \qquad \text{(EC.70)}$$

$$= \frac{\mu + r}{1 + r} \qquad \text{(EC.71)}$$

Next, consider a selection algorithm without access to v. Let p_0 be the fraction of candidates selected by the algorithm that have $\mathbf{x}$ *2* W_0. Since the algorithm has no access to v, the fraction of selected candidates who are majorities is

$$= \ 2sr + p_0 \cdot (1 + r)$$

$$s_0 = p_0 \cdot \mathrm{Pr} fv = 0 j\mathbf{x}\ 2\ \mathrm{W}_0 g + (1 - p_0) \cdot \mathrm{Pr} fv = 0 j\mathbf{x}\ 2\ \mathrm{W}_1 g \qquad \text{(EC.72)}$$

$$= (1 - p_0) \cdot \frac{1 + r}{2sr + 1 + r}; \qquad \text{(EC.73)}$$

while the fraction of minorities is

$$s_1 = p_0 \cdot \mathrm{Pr} fv = 1 j\mathbf{x}\ 2\ \mathrm{W}_0 g + (1 - p_0) \cdot \mathrm{Pr} fv = 1 j\mathbf{x}\ 2\ \mathrm{W}_1 g \qquad \text{(EC.74)}$$

$$= p_0 + (1 - p_0) \cdot \frac{2sr}{2sr + 1 + r}: \qquad \text{(EC.75)}$$

Gi$fv = 0g$ .e.$\mathrm{Pr} fv = 1g$ the $s_1 = s_0 \geq r$ requirementn of AIR

$(1 - p_0) \cdot (1 + r)$

$(1 + p_0^2) \cdot (1 + r)$

Recall that the expected quality of candidates with $\mathbf{x}\ 2\ \mathrm{W}_0$ and $\mathbf{x}\ 2\ \mathrm{W}_1$ is μ and μ_1, respectively, with $\mu \leq \mu_1$. This leads to the following upper bound on p_{SE}:

$$\mathrm{p}_{\mathrm{SE}} \leq \mu \cdot \frac{r + r^2 - 2sr}{(1+r)^2} + \mu_1 \cdot \quad 1 - \frac{r + r^2 - 2sr}{(1+r)^2} \tag{EC.79}$$

$$= \mu \cdot \frac{r + r^2 - 2sr}{(1+r)^2} + \frac{2sr + \mu(1+r)}{2sr + 1 + r} \cdot \quad 1 - \frac{r + r^2 - 2sr}{(1+r)^2} \tag{EC.80}$$

$$= \mu \cdot \frac{r + r^2 - 2sr}{(1+r)^2} + \frac{2sr + \mu(1+r)}{2sr + 1 + r} \cdot \frac{1 + r + 2sr}{(1+r)^2} \tag{EC.81}$$

$$= \mu \cdot \frac{r + r^2 - 2sr}{(1+r)^2} + \frac{2sr + \mu(1+r)}{(1+r)^2} \tag{EC.82}$$

$$= \frac{\mu(1+r)^2 + 2sr(1-\mu)}{(1+r)^2} \tag{EC.83}$$

Putting together p_{SE} and $\mathrm{p}_{\max}$, we have

$$\frac{\mathrm{p}_{\mathrm{SE}}}{\mathrm{p}_{\max}} \leq \frac{\mu \cdot (r + r^2 - 2sr) + 2sr + \mu(1+r)}{(1+r)r + (1+r)\mu}: \tag{EC.84}$$

Since $s\ 2\ (0; 1{=}2]$, there is $r + r^2 - 2sr \geq r^2 \geq$