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 race or gender. Due to prevailing laws in the United States (i.e., Civil Rights Act 1991) and the European Union (i.e., Race Equality Directive 2000/43/EC), ML algorithms are *not* allowed to directly access the race or gender of an applicant (so as to avoid the so-called "disparate treatement"; Gottfredson 1994). Yet, as ML researchers soon realized, many other predictors (e.g., test scores for cognitive ability; Hunter and Hunter 1984) contain information about the protected demographic variables, which could be learned by ML algorithms to generate unfair outcomes. As a case in point, Amazon developed an ML algorithm for screening resumes, only to scrap it altogether because the algorithm learned rules (from historic data) that automatically penalize candidates who graduated from women's colleges (Dastin 2018). Similar biases were found for ML algorithms in other selection settings, such as loan underwriting (Fu et al. 2021, Martinez and Kirchner 2021), tenant screening (Gikay 2020), college admission (Burke 2020), etc. To address this crucial concern in designing ML algorithms, researchers started pursuing certain *fairness goals* over the selection outcome generated by the ML algorithm. For example, a common fairness goal for personnel selection is a lower bound on the adverse-impact ratio (AIR), which is the ratio between selection rates for minority and majority candidates (Zhang et al. 2023). ML algorithms that pursue at least one fairness goal are collectively known as *Fairness-Aware ML* (FAML) algorithms, which have seen important contributions from multiple related fields such as operations management (De-Arteaga et al. 2022, Kallus et al. 2022), information systems (Fu et al. 2022), and computer science (Mehrabi et al. 2021).

Nearly all existing FAML algorithms assume that the selection outcome is directly derived from the prediction model generated by the ML algorithm (cf. Mehrabi et al. 2021). In other words, when a firm employs the prediction model to calculate the predicted scores for all applications, it proceeds to select (i.e., approve) those applications with predicted scores above a certain threshold[1]. In practice, however, ML predictions are seldom used as-is in selection settings. Take personnel selection, for example, where ML typically serves a supportive role to Human Resource (HR) managers by screening candidates for subsequent interviews (Liem et al. 2018). That is, ML predictions are primarily utilized to identify and screen out unqualified cases, whereas for the remaining applicants, HR managers manually gather additional information (e.g., through interviews and reference letters) before making decisions. Consequently, the task assigned to ML is more appropriately categorized as a *screening task* rather than a selection task in practice.

Despite this distinction, the prevailing view today is that it has little impact on the design and functioning of FAML algorithms, except for the obvious (quantitative) difference in selection/retention rate. The prevalence of this view is evidenced by the near-universal adoption of the selection task in defining optimization goals for ML algorithms (Mehrabi et al. 2021). On the surface, this view also seems reasonable, as the goal of ML is to assign a "better" application with a higher predicted score. Whether one application is deemed "better" than another seems irrelevant to whether the task at hand is selection or screening.

[1] The threshold could be pre-determined or calculated based on the top-k scores (e.g., if the number of approvals is limited by k).

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] – e.g., either very good or very bad. Opting for such an applicant is inherently risky for the selection task, given the prospect of hiring an unqualified employee. For the screening task, however, the dynamics shift. Even in scenarios where the applicant turns out to be unqualified, a capable HR manager still has the chance to catch and reject the application after an interview, thereby limiting the cost of this risky choice to a wasted interview spot. In other words, an FAML algorithm should logically lean towards making high-risk high-reward choices if it were equipped with the knowledge of (the existence of) latter-stage interviews. Conversely, to design an FAML algorithm for selection, only to subsequently employ it for screening, misses this opportunity afforded by the screening task – an opportunity to leverage the availability of latter-stage interviews to manage the consequences of bold choices. Further, we demonstrate that the ability to make high-risk high-reward choices is crucial for ameliorating the fairness problems known to exist for FAML algorithms, such as selection biases within the minority group and the resulting between-group differences in selection quality (Zhang et al. 2023). This highlights the importance of designing FAML algorithms for the screening task rather than mis-categorizing a screening task as a selection one in algorithmic design.

The rest of the paper is organized as follows. We briefly review the related literature in Section 2. Then, we start by comparing ML selection and screening algorithms without fairness constraint in Section 3, followed by a comparison with fairness constraint in Section 4. Section 5 presents mathematical analysis and experimental results (over both simulated and real-world data) on the comparison between selection and screening. We conclude the paper with discussions of its managerial implications and limitations in Section 6.

2 Literature Review

Organizational selection decisions, such as personnel selection, have been extensively studied across scientific disciplines such as operations management (e.g., Aksin et al. 2007), human resource management (SIOP 2018), machine learning (e.g., Mitchell et al. 2018), etc. In the operations management literature, for example, personnel selection serves as the foundation of the renowned *secretary problem*, which has continued to garner attention over the past half-century (e.g., Lindley 1961, Stewart 1981, Tamaki 1991, Eriksson et al. 2007, Oh and Özer 2016, Kesselheim et al. 2023). More broadly, operations management scholars have delved into various aspects of personnel selection, exploring the link between hiring decisions and employee learning (Arlotto et al. 2014), investigating the complex interactions among hiring, training,

[2] As explained later in the paper, the optimization goals pursued by FAML algorithms make such bimodal distributions common occurrences for FAML predictions.

and turnovers (Aksin et al. 2007), and examining the interface between operations management and human resource management (Boudreau et al. 2003). As our paper focuses on the use of FAML algorithms in selection settings, this literature review zeros in on two issues specific to the use of FAML: 1) what "fairness" means in organizational selection settings; and 2) the existing design of FAML algorithms.

2.1 Fairness Requirements

In terms of what "fairness" means in organizational settings such as hiring and promotions, the prevailing view in both research and practice is that "fairness has no single meaning and, therefore, no single definition" (SIOP 2018). As one cannot exhaust all ideological definitions of fairness, a focus of the existing FAML literature is on satisfying the *legal* mandates that apply to real-world selection settings, especially in the context of the US legal system. In the passages that follow, we review the two main legal mandates, the ban of *disparate treatment* and *disparate impact*, respectively. Note that these legal mandates require a firm to eliminate *both* disparate treatment and disparate impact when making selection decisions.

2.1.1 Disparate Treatment

Disparate-treatment laws prohibit the use of legally protected variables, such as race, ethnicity, gender, national origin, etc., in making selection decisions (Gottfredson 1994, Primus 2010). For example, the US Civil Rights Act (CRA) of 1968 prohibits the use of such protected variables in making lending decisions. Similarly, both the US CRA of 1991 and the European Union Race Equality Directive 2000/43/EC stipulate the same ban in the employment context. The US CRA 1991 explicitly outlaws a then-existing practice of subgroup-norming, making it illegal to "alter the results of employment related tests on the basis of race, color, religion, sex, or national origin" (US CRA 1991, §106).

Enforcing disparate-treatment laws is clearly challenging because many protected variables are readily accessible in selection settings. For example, in job interviews, the protected variables of an interviewee are often directly observable by interviewers, potentially triggering conscious or unconscious biases in their decision making (Purkiss et al. 2006). To prevent such disparate treatment from happening in algorithmic selections, a general consensus in the FAML literature is that FAML algorithms should *not* be allowed to access any protected variable of an applicant (Kroll et al. 2016). In other words, the algorithm should *not* be given information about whether an applicant is in the minority or majority group. We follow this rule throughout the paper.

It is important to understand that banning algorithmic access to protected variables is a necessary but *insufficient* condition for making "fair" selection decisions. A key reason here is that other predictor variables, from location (e.g., ZIP code) to test scores, could well contain information that serves as proxies for protected variables (Pedreshi et al. 2008). In the aforementioned Amazon example, even though the algorithm did not overtly access the gender of any applicant, it did have access to applicants' education background (e.g., attendance in women's colleges), which became a proxy for the protected gender variable. Disparate-impact laws, which we review next, were developed to prevent such "covert" disparity manifested by facially neutral practices.

2.1.2 Disparate Impact

Unlike disparate-treatment laws which regulate the input to an organizational decision process, disparate-impact laws regulate the output of it. Specifically, disparate-impact laws stipulate that there should *not* be a gross statistical disparity in the selection outcome for minority and majority candidates *unless* such disparate impact can be demonstrably justified by a significant, legitimate, business necessity (National Research Council 2004). Further, in the US legal system, a firm may be held liable if there exists an alternative selection process that results in less statistical disparity while serving the firm's legitimate business needs (42 U.S.C. §2000e–2). This creates a clear incentive for firms to reduce any statistical disparity in a selection process (Oswald et al. 2016).

In terms of how to measure the statistical disparity, one of the most widely used metric in practice is the *adverse-impact ratio* (AIR) proposed by the US Equal Employment Opportunity Commission (EEOC) in the *Uniform Guidelines on Employee Selection Procedures* (29 C.F.R. §1607, 1978). Specifically, AIR is defined as the ratio between the selection rate for minority candidates and that for the majorities. Given the wide adoption of AIR in research and practice (De Corte et al. 2011), we use it as the measure of disparate impact in this paper. The larger the AIR, the less disparate impact there is in the selection outcome. Without introducing ambiguity, we use "fairness constraint" to refer to the requirement of AIR $\geq r$, where $r \in [0, 1]$ is a pre-determined threshold, throughout the paper.

2.2 FAML Algorithms

In general, an FAML algorithm uses a training dataset – typically consisting of data from past selection decisions – to generate a prediction model for a certain *prediction target* specified to the FAML algorithm. As elaborated in latter sections, the focus of our paper is on how the prediction target should be specified for an FAML algorithm, and whether the specification should differ for selection and screening tasks. The technical design of ML algorithms used to (train a prediction model to) approximate the prediction target, including the functional form of the prediction model, is an issue orthogonal to the focus of this paper. Thus, we keep the review of FAML algorithms brief in the rest of this section.

The FAML literature includes a wide variety of designs, from linear models (Berk et al. 2018) to neural networks (Louizos et al. 2016), from Gaussian processes (Tan et al. 2020) to support-vector machines (Zafar et al. 2019). Many of these designs are mathematically guaranteed to asymptotically converge to the Bayes error when the size of the input training dataset increases (e.g., Gaussian process; Rasmussen and Williams 2006). Since our focus is on the distinction between selection and screening tasks in the prediction target, 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 constraint. We first analyze the optimal design of ML algorithm for selection and screening, respectively. Then, we present an illustrative example to demonstrate the difference in outcome between the two algorithms when both are used to retain the same number of candidates for manual interviews.

3.1 ML for Selection Task

Population of Candidates: Consider a selection setting in which each candidate (e.g., a loan application, a job candidate) is described by $\langle \mathbf{x}, v \rangle$, where $\mathbf{x} \in \Omega$ is a vector of characteristics that can be observed in the selection process (with Ω representing the domain of $\mathbf{x}$), and $v \in \{0, 1\}$ indicates whether the candidate belongs to a protected group (e.g., of racial/ethnic minorities). We refer to individuals with $v = 1$ as the protected *minorities*, and those with $v = 0$ as the majority group. Each candidate is also associated with a non-negative real-valued variable $y \in [0, \infty)$, which represents the candidate's quality of interest. In personnel selection, for example, $\mathbf{x}$ would contain characteristics revealed by a job application, such as an individual's cognitive ability, personality, biodata (i.e., past experience), etc. y would represent the future job 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 Γ 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* as a function of $\mathbf{x}$, namely $L(\mathbf{x}) \in [0,1]$, which describes the probability for a candidate with characteristics $\mathbf{x}$ to be selected. Note that this notation captures the more general setting of stochastic selection decisions. In the special case where ML makes deterministic decisions for a given $\mathbf{x}$, $L(\mathbf{x})$ would simply be limited to one of the two extreme values 0 or 1.

In real-world selection scenarios such as personnel selection, there is usually a pre-determined limit on the fraction of candidates that can be admitted. To capture this limit, we assume the selection decisions to be capacity-constrained, meaning that

$$\int_{\Omega} L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s, \tag{1}$$

where $p_{\Gamma}(\mathbf{x})$ is the marginal probability density function of $\mathbf{x}$ according to the joint distribution Γ, and $s \in [0,1]$ is the *selection rate*, i.e., the maximum fraction of candidates that can be selected.

Subject to the capacity constraint, the ML algorithm is designed to maximize the expected[3] quality of selected candidates. This is equivalent with finding an optimal selection decision function L^* that satisfies

$$\begin{aligned} L^* = \arg\max_{L} \int_{\Omega} \mathbb{E}_{\Gamma}(y|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \\ s.t. \int_{\Omega} L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s, \end{aligned} \tag{2}$$

where $\mathbb{E}_{\Gamma}(y|\mathbf{x})$ is the conditional expectation of y given $\mathbf{x}$ according to the joint distribution Γ.

Optimal ML Design: Rambachan et al. (2020, Proposition 1) proved[4] that the optimal solution to Equation 2 has a strikingly simple form. That is, for any given Γ, there always exists a constant $c \geq 0$, such that $L^*(\mathbf{x}) = 1$ if $\mathbb{E}_{\Gamma}(y|\mathbf{x}) \geq c$ and $L^*(\mathbf{x}) = 0$ otherwise[5]. In other words, if the ML algorithm can accurately estimate the expected quality $\mathbb{E}_{\Gamma}(y|\mathbf{x})$ of each candidate, then its optimal choice would be to simply admit candidates in a decreasing order of their expected quality (i.e., from high to low) until reaching the capacity constraint. Whereas the mathematical proof is subtle, the conclusion is surprisingly consistent with what has been pursued in the field of personnel selection for decades, i.e., to maximize the expected quality (also known as *expected criterion*) of candidates being selected (De Corte et al. 2011).

[3] Note that expectation is taken over the randomness of both the latent quality variable y and the stochastic selection decisions $L(\cdot)$.

[4] At a high level, the proof can be considered an extension of the classic Hardy-Littlewood inequality (Hardy et al. 1952).

[5] This statement assumes a finite density function for the marginal distribution of y according to Γ. If the assumption is violated, then we might need a tie-breaking design of $L^*(\mathbf{x}) \in [0,1]$ when $\mathbb{E}_{\Gamma}(y|\mathbf{x}) = c$ to satisfy the capacity constraint. See Rambachan et al.'s (2020) proof for details.

With this, the ML algorithm design is then reduced to generating an accurate point estimate of $\mathbb{E}_\Gamma(y|\mathbf{x})$ for a given $\mathbf{x}$. To do so, the ML algorithm learns from a training dataset formed by historic instances of $\langle \mathbf{x}, v, y \rangle$ which are assumed to be drawn from the same joint distribution Γ. 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 $\langle \mathbf{x}, v, y \rangle$ to be sufficiently large so as to allow ML to learn the joint distribution Γ 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 $\langle \mathbf{x}, v, y \rangle$ drawn from joint distribution Γ, the ML algorithm only has access to the candidate's characteristics vector $\mathbf{x}$, and admits the candidate with probability $L(\mathbf{x}) \in [0,1]$. Unlike in the selection setting, the candidates admitted by the ML algorithm are not directly selected. Instead, they enter a second-stage interview process in which the final selection decisions are made by human experts (e.g., HR managers). This change has two main implications in the design of ML algorithm.

First, an analysis of the final selection quality now requires a model of the manual interview process. In practice, human experts may gather additional information beyond the characteristics in $\mathbf{x}$ during the interview process. They may also opt to collect different information for different candidates. Such flexibility makes the *process* of manual interview extremely difficult to model. To address the challenge, we instead model the *outcome* of the interview process denoted by a binary variable $\mathbb{T} \in \{0,1\}$, with $\mathbb{T} = 1$ representing a candidate passing the interview and $\mathbb{T} = 0$ otherwise.

Since the purpose of this paper is to examine the design of ML algorithms for screening, we consider manual interviews that yield the best possible selection outcome given the ML screening results. In other words, the manual interview selects an optimal subset (in terms of expected quality) of candidates who passed ML screening (subject to the selection rate constraint). Obviously, the optimal subset is formed by candidates with quality y over a certain cutoff y_0. That is, $\mathbb{T} = \mathbb{1}(y \geq y_0)$, where $\mathbb{1}(y \geq y_0)$ is the indicator function that returns 1 if $y \geq y_0$ and 0 otherwise. With this, the pursuit of selection quality is equivalent with maximizing

$$u = \int_\Omega \mathbb{E}_\Gamma(y \cdot \mathbb{T}|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \tag{3}$$

$$= \int_{\Omega} \mathbb{E}_{\Gamma}(y \cdot \mathbb{1}(y \geq y_0)|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \tag{4}$$

$$= \int_{\Omega} \mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x}) \cdot \Pr\{y \geq y_0|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x}, \tag{5}$$

where $\mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x})$ is the expectation of y given $\mathbf{x}$ conditional upon $y \geq y_0$. Intuitively, Equation 5 suggests that, in the screening case, only those candidates who can pass the manual interview matters for the final selection quality.

Second, the introduction of manual interviews also alters the capacity constraints. There are now two such constraints governing ML screening and final selection, respectively. For ML screening, there must be

$$\int_{\Omega} L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s_1, \tag{6}$$

where s_1 is the maximum fraction of candidates that can be retained to manual interviews. Then, the final selection must obey

$$\int_{\Omega} \Pr\{y \geq y_0|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s, \tag{7}$$

where s is the final selection rate.

Putting together Equation 5 with the two capacity constraints, we see that the objective of the ML algorithm under the screening setting is to find

$$L^* = \arg\max_{L} \int_{\Omega} \mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x}) \cdot \Pr\{y \geq y_0|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x}$$

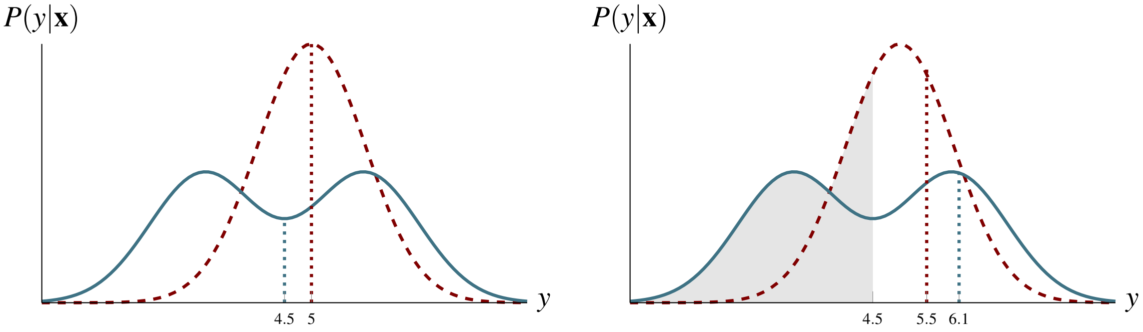
$$s.t. \int_{\Omega} L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s_1 \text{ and } \int_{\Omega} \Pr\{y \geq y_0|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s. \tag{8}$$

Optimal ML Design: With two capacity constraints, the optimization problem becomes considerably more difficult to solve compared with the selection case. To ease the discussions, we start with a simplifying assumption that the interview cost is low – i.e., s_1 is sufficiently large so as to make Inequality 7 the only capacity constraint that matters – before removing this assumption later in mathematical and experimental analyses. Note that this simplification does not imply a trivial ML solution of retaining all candidates (i.e., setting $L^*(\mathbf{x}) = 1$ for all $\mathbf{x}$) because doing so may violate the capacity constraint on final selections (i.e., Inequality 7). With the simplification, Rambachan et al.'s (2020) proof directly carries over to the screening case, with the only change (from the selection case) being the replacement of $\mathbb{E}_{\Gamma}(y|\mathbf{x})$ with $\mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x})$. In other words, under the screening setting, the optimal choice for the ML algorithm is to retain candidates with characteristics $\mathbf{x}$ in a decreasing order of $\mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x})$ – i.e., their expected quality *conditional upon passing the manual interview* – until reaching the capacity constraint.

Compared with the selection case, the difference is straightforward. In the selection case, the expected quality of every ML-admitted candidate matters (hence the ranking of $\mathbb{E}_{\Gamma}(y|\mathbf{x})$), because they all affect the final selection quality. In the screening case, however, only those ML-admitted candidates who can pass the manual interview matters (hence the ranking of $\mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x})$), because the others affect only interview costs but not the final selection quality. As elaborated next, this key difference gives a screening-oriented ML algorithm the ability to make high-risk high-reward choices in retaining candidates.

3.3 Comparison between Selection and Screening

Whereas the previous two subsections explicate the design difference of ML algorithms for selection and screening, we now examine how such design differences lead to different outcomes when both algorithms are used in the exact 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. For each candidate, both algorithms have access to the exact same information, i.e., characteristics $\mathbf{x}$ and nothing else. Both algorithms are also given the same training dataset formed by historic instances of $\langle \mathbf{x}, v, y \rangle$. The selection algorithm is simply specified to choose s_1 fraction of candidates. The screening algorithm is specified to do the same, but is also given s as input, with the understanding that s out of the s_1 retained candidates will be eventually selected. This way, the screening algorithm can compute the quality cutoff y_0 accordingly.



(a) Selection

(b) Screening ($y_0 = 4.5$)

Figure 1 Illustrative Example for Selection vs. Screening without Fairness Constraint

Note. Both panels depict the probability density function of $P(y|\mathbf{x})$ for Alice (red dashed) and Bob (blue solid). Panel (a) shows that Alice has a higher expected quality (i.e., $\mathbb{E}_\Gamma(y|\mathbf{x}) = 5$) than Bob (4.5), and is thus the preferred choice in the selection setting. For the screening setting, Panel (b) shows, when we consider the expected quality conditional upon passing a subsequent interview with $y_0 = 4.5$, then Bob becomes the preferred choice with $\mathbb{E}_\Gamma(y|y \geq y_0, \mathbf{x}) = 6.1$ over Alice (5.5). The shaded region in Panel (b) represents scenarios where Alice or Bob is filtered out by the manual interview.

To highlight differences between the two algorithms, we consider an illustrative example comparing two candidates Alice and Bob, and explain why the two algorithms would have different preferences between them. Recall from earlier discussions that both algorithms make their decisions about a candidate based entirely on the conditional probability distribution of quality y given the candidate's characteristics $\mathbf{x}$ (i.e., $P(y|\mathbf{x})$), which we refer to as the *quality distribution* of a candidate. In the example, we consider quality distributions for Alice and Bob as depicted in Figure 1. For Alice, the distribution is Gaussian $N(5,1)$, i.e., with mean 5 and variance 1. For Bob, the distribution is a Gaussian mixture with two components[6] of equal weight, one being $N(3,1)$ and the other being $N(6,1)$. As can be seen from the figure, Bob's quality

[6] As elaborated in the next section, such bimodal Gaussian mixture is indeed common occurrence in quality distribution for FAML.

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 $\mathbb{E}_\Gamma(y|\mathbf{x})$ 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., $\mathbb{E}_\Gamma(y|y \geq y_0, \mathbf{x})$), 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 be filtered out by the interview anyway, affecting only the interview cost yet having zero effect on the final selection quality. This de-risking feature of manual interview is what allows the screening algorithm to make high-risk high-reward choices (like Bob) that the selection algorithm is unable to make.

In sum, even when both selection and screening algorithms are used in the same way (i.e., to retain s_1 fraction of candidates), they could reach different conclusions on whether one candidate is "better" than another, and therefore produce different outcomes. As elaborated in the next section, this difference is a key contributor to the known fairness issues incurred by FAML algorithms designed for the selection setting.

4 Fairness Implications of Selection vs. Screening

We now examine the differences between ML for selection and screening *with* the presence of fairness constraint. Like in the last section, we first analyze the optimal design of FAML algorithms for selection and screening, respectively, before using an illustrative example to explain their difference when both are used to retain candidates for manual interviews. Further, we show that this difference directly addresses the known fairness issues incurred by FAML algorithms designed for the selection setting.

4.1 Selection with Fairness Constraint

Recall from Section 2.1.2 that we focus on fairness constraint expressed as a lower bound on the adverse impact ratio (AIR), i.e., AIR $\geq r$, where AIR is the ratio between the selection rates of minorities and majorities. For example, a constraint of AIR ≥ 1 would require the selection rate for minorities to be at least as high as that for majorities. Given the fraction of minority candidates p_1 (according to the joint distribution Γ) and the selection rate s, simple algebraic transformations can reduce AIR $\geq r$ to

$$\int_\Omega \Pr\{v = 1|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \geq \frac{r \cdot s \cdot p_1}{1 - p_1 + p_1 \cdot r}, \tag{9}$$

where $\Pr\{v=1|\mathbf{x}\}$ is the conditional probability of a candidate being a minority given $\mathbf{x}$ according to Γ, because otherwise there would be

$$\text{AIR} = \frac{\frac{1}{p_1}\cdot\int_\Omega \Pr\{v=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x}}{\frac{1}{1-p_1}\cdot\left(s-\int_\Omega \Pr\{v=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x}\right)} < \frac{\frac{r\cdot s}{1-p_1+p_1\cdot r}}{\frac{1}{1-p_1}\cdot\frac{s-s\cdot p_1}{1-p_1+p_1\cdot r}} = r. \tag{10}$$

Putting together this new AIR constraint with Equation 2, the objective of an FAML algorithm (i.e., with fairness constraint) becomes to find

$$\begin{aligned}
&L^* = \arg\max_L \int_\Omega \mathbb{E}_\Gamma(y|\mathbf{x})\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x}\\
&s.t. \int_\Omega L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s \text{ and } \int_\Omega \Pr\{v=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \ge \frac{r\cdot s\cdot p_1}{1-p_1+p_1\cdot r}
\end{aligned} \tag{11}$$

Zhang et al. (2023, Theorem 1) proved that the optimal solution to Equation 11 can be deduced through the method of Lagrange multiplier (Nocedal and Wright 2006). That is, for any given s, p_1, and r, there always exists a Lagrange multiplier $\lambda \ge 0$, such that Equation 11 is equivalent with

$$\begin{aligned}
&L^* = \arg\max_L \int_\Omega \left(\mathbb{E}_\Gamma(y|\mathbf{x}) + \lambda\cdot\Pr\{v=1|\mathbf{x}\}\right)\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x}\\
&s.t. \int_\Omega L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s.
\end{aligned} \tag{12}$$

Clearly, the larger r is, the larger λ will be. When $r=0$, we have $\lambda=0$, reducing Equation 12 to the baseline selection case without fairness constraint.

With this transformation, Rambachan et al.'s (2020) Proposition 1 can again be directly applied. That is, under the selection setting with fairness constraint, the optimal choice for the FAML algorithm is to admit candidates with characteristics $\mathbf{x}$ in a decreasing order of

$$f(\mathbf{x}) = \mathbb{E}_\Gamma(y|\mathbf{x}) + \lambda\cdot\Pr\{v=1|\mathbf{x}\} \tag{13}$$

until reaching the capacity constraint. Compared with the selection case without fairness constraint, Equation 13 includes a new additive component of $\lambda\cdot\Pr\{v=1|\mathbf{x}\}$. In other words, instead of predicting quality alone, the FAML algorithm is designed to predict a weighted sum of the expected quality and the probability for the candidate to be a minority.

4.2 Screening with Fairness Constraint

A similar transformation can be carried out for the screening task with fairness constraint. Specifically, putting together the selection quality in Equation 3 with the fairness constraint AIR $\ge r$, we see that the objective of an FAML screening algorithm is to find

$$\begin{aligned}
&L^* = \arg\max_L \int_\Omega \mathbb{E}_\Gamma(y\cdot\mathbb{T}|\mathbf{x})\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x}\\
&s.t. \int_\Omega L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s_1, \int_\Omega \Pr\{\mathbb{T}=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s,\\
&\text{and } \int_\Omega \Pr\{v=1,\mathbb{T}=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \ge \frac{r\cdot s\cdot p_1}{1-p_1+p_1\cdot r},
\end{aligned} \tag{14}$$

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

$$\begin{aligned}
L^* &= \arg\max_L \int_\Omega (\mathbb{E}_\Gamma(y \cdot \mathbb{T}|\mathbf{x}) + \lambda \cdot \Pr\{v=1, \mathbb{T}=1|\mathbf{x}\}) \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \\
&= \arg\max_L \int_\Omega (\mathbb{E}_\Gamma(y|\mathbb{T}=1,\mathbf{x}) \cdot \Pr\{\mathbb{T}=1|\mathbf{x}\} + \lambda \cdot \Pr\{v=1|\mathbb{T}=1,\mathbf{x}\} \cdot \Pr\{\mathbb{T}=1|\mathbf{x}\}) \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \\
&= \arg\max_L \int_\Omega (\mathbb{E}_\Gamma(y|\mathbb{T}=1,\mathbf{x}) + \lambda \cdot \Pr\{v=1|\mathbb{T}=1,\mathbf{x}\}) \cdot \Pr\{\mathbb{T}=1|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \\
s.t. & \int_\Omega \Pr\{\mathbb{T}=1|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \le s, \qquad (15)
\end{aligned}$$

where λ ($\lambda \ge 0$) 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

$$f'(\mathbf{x}) = \mathbb{E}_\Gamma(y|\mathbb{T}=1,\mathbf{x}) + \lambda \cdot \Pr\{v=1|\mathbb{T}=1,\mathbf{x}\} \qquad (16)$$

until reaching the capacity constraint.

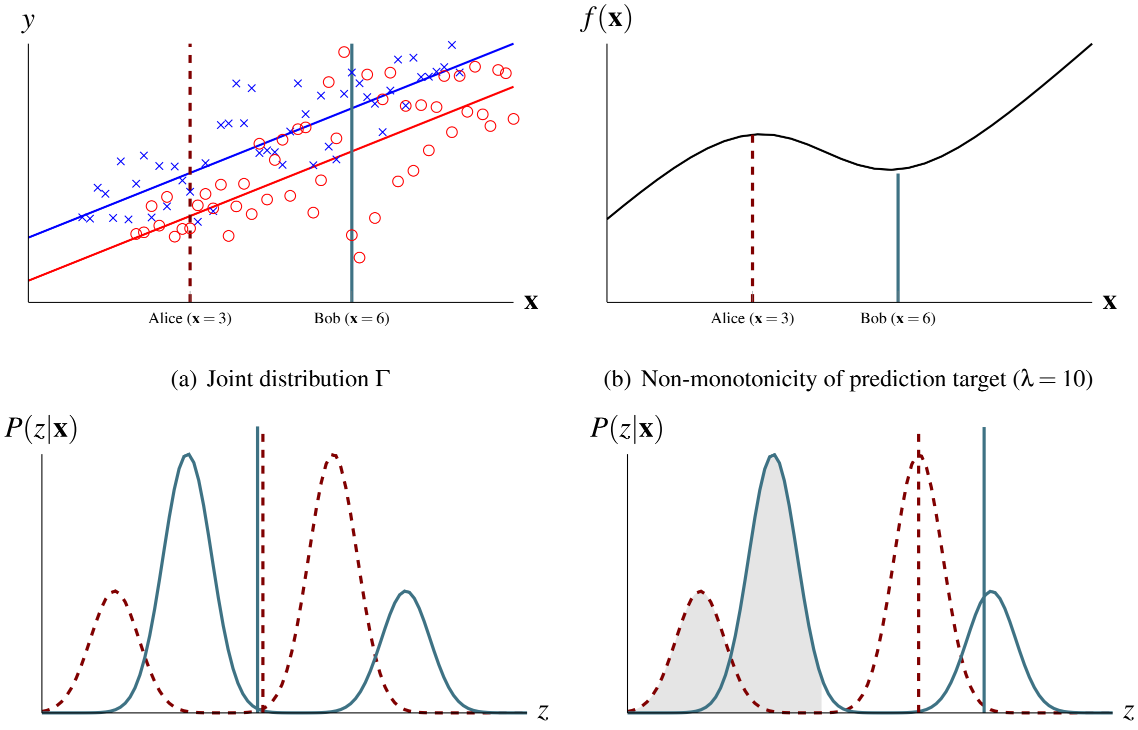
Juxtaposing Equation 16 with the optimal design for the selection case (i.e., Equation 13), the difference is, in essence, the same as the selection-screening difference without fairness constraint. That is, for screening, only candidates who can pass the manual interview matters for final selection 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 $\mathbb{T} = \mathbb{1}(y + \lambda_2 v \ge t_0)$, 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

$$f'(\mathbf{x}) = \mathbb{E}_\Gamma(y|y+\lambda_2 v \ge t_0, \mathbf{x}) + \lambda \cdot \Pr\{v=1|y+\lambda_2 v \ge t_0, \mathbf{x}\} \qquad (17)$$

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 ($s \le s_1$) fraction of candidates who must satisfy the fairness constraint of AIR $\ge r$. 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



(a) Joint distribution Γ

(b) Non-monotonicity of prediction target ($\lambda = 10$)

(c) Selection ($z = y + \lambda \cdot v$, $\lambda = 10$)

(d) Screening ($\lambda = \lambda_2 = 10, t_0 = 8$)

Figure 2 Illustrative Example for Selection vs. Screening with Fairness Constraint.

Note. Panel (a) depicts the joint distribution Γ, with × and ○ representing minorities and majorities, respectively. Panel (b) shows that $f(\mathbf{x})$, the prediction target for FAML selection algorithm, does not monotically increase with $\mathbf{x}$. Panels (c) and (d) depict the probability density function of $P(z|\mathbf{x})$ for Alice (red dashed) and Bob (blue solid), where $z = y + \lambda \cdot v$. The two vertical lines in Panel (c) show the expected value of z (i.e., the prediction target for selection algorithm) for Alice and Bob, while those in Panel (d) show the conditional expectation of z given $z \geq t_0$ (i.e., the prediction target for screening algorithm). The shaded region in Panel (d) represents scenarios where Alice or Bob is filtered out by the manual interview (i.e., $z < t_0$).

interview, it has to be specified to choose s_1 fraction of candidates while satisfying the fairness constraint (i.e., AIR $\geq r$) for the chosen candidates (as is done in almost all existing work; Mehrabi et al. 2021). The screening algorithm, thanks to it accounting for the manual interview, is specified to choose s_1 fraction of candidates, but is also given the final selection rate s as input with the requirement that the final selection outcome satisfying AIR $\geq r$.

To highlight differences between the two algorithms, we once again consider their preferences between two candidates, Alice and Bob. To make the comparison more realistic, we no longer use hypothetical quality distributions as in the earlier example (Section 3.3). Instead, we start with defining a simple joint distribution Γ according to empirical findings in the personnel selection literature. Specifically, ample empirical evidence suggests that, compared with a majority candidate with the same quality, a minority candidate tends to score lower in predictors that are positively correlated with quality (e.g., SAT scores for college admission, cognitive ability tests in personnel selection; Berry et al. 2011), with a possible reason being that 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,

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

where $\varepsilon \sim N(0,1)$ 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 *de-risks* the selection of Bob because his left (i.e., “risky”) component is now mostly filtered out by the manual interview. As a result, the screening algorithm no longer needs to skip a higher-quality candidate (i.e., Bob) to choose a low-risk alternative (i.e., Alice). Indicatively, the screening prediction target for Bob (14.70) is now larger than Alice (12.00), meaning that the FAML algorithm for screening prefers Bob over Alice, consistent with the order of their expected quality. As can be seen from this example, it is the screening algorithm’s ability to make high-risk high-reward choices that ameliorates the within-group selection bias of FAML selection algorithms. In other words, it is crucial to properly specify the screening task to an FAML algorithm rather than miscategorizing it as a selection task.

5 Mathematical Analysis, Simulation, and Experimental Studies

Conceptual development in the previous two sections reveals a difference between FAML selection and screening algorithms in their risk-taking tendency and, consequently, their preferences between different candidates. This finding raises three questions to be answered through mathematical analysis and experimental studies. The first question is on whether within-group selection bias (depicted in Figure 2b) always manifests as quality degradation when FAML algorithms are directly used in a selection setting. Answering this question could help determine whether it is necessary to incur the cost of manual interviews in practice. We use mathematical analysis to study this question in the first part of this section.

Then, the second question is on whether assigning FAML with the screening task followed by manual interviews could ameliorate the quality degradation caused by within-group selection bias. Answering this question could help us understand whether addressing the within-group selection bias of FAML selection algorithms requires a change of the current legal system (as argued in existing work, e.g., Lipton et al. 2018), or if such bias could be alleviated by simply correcting the mis-categorization of screening task as a selection one. We examine this second question through simulation studies in the second part of this section.

Finally, once we establish the superiority of using FAML followed by manual interviews, the third question that arises is how the actual performance of FAML selection and screening algorithms differ when *both* are used to screen candidates for manual interviews. Since this comparison depends on not only the data distribution but the ML algorithm being used, we address it through experimental studies on both simulated and real-world data in the last part of this section.

5.1 Mathematical Analysis

A fairness constraint is only applicable when the distributions of predictors $\mathbf{x}$ 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 Γ. Specifically, we are interested in between-group difference on $P(y|\mathbf{x})$, the conditional distribution of y given $\mathbf{x}$, because FAML selection algorithm relies on $P(y|\mathbf{x})$ in their decision-making. To capture between-group difference on $P(y|\mathbf{x})$, 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 Γ is defined as

$$\delta_\Gamma = \max_{\Theta \subseteq \Omega} \left| \frac{\mathbb{E}_\Gamma(y|\mathbf{x} \in \Theta, v=0) - \mathbb{E}_\Gamma(y|\mathbf{x} \in \Theta, v=1)}{\mathrm{SD}_\Gamma(y|\mathbf{x} \in \Theta)} \right|, \tag{19}$$

where Ω is the domain of $\mathbf{x}$, $|\cdot|$ is the absolute value, and SD_Γ represents standard deviation over Γ.

In terms of the value of δ_Γ in practice, Roth et al. (2003) show that between-group difference varies depending on the quality measure being used, e.g., from 0.13 for a subjective measure of absenteeism to 0.52 for an objective measure of work samples to 0.55 for an objective measure of job knowledge.

Besides δ_Γ, we also need a way to detect the within-group selection bias discussed in Section 4.3. Recall that such a bias manifests as a reduction of selection quality because, within the minority (or majority) group, it could favor a candidate with lower expected quality over a higher-quality candidate. Thus, we detect the presence of within-group selection bias by comparing the final selection quality of FAML algorithms with the maximum possible selection quality subject to capacity (i.e., selection rate s) and fairness (i.e., AIR $\geq r$) constraints. As discussed in Section 4.2, this ideal selection quality can be captured by

$$\begin{aligned}
\pi_{\max} = \max_{t_0, t_1} \Bigg(& \int_\Omega \mathbb{E}_\Gamma(y|y \geq t_1, v=1, \mathbf{x}) \cdot \Pr\{y \geq t_1 | v=1, \mathbf{x}\} \cdot \Pr\{v=1|\mathbf{x}\} \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} + \\
& \int_\Omega \mathbb{E}_\Gamma(y|y \geq t_0, v=0, \mathbf{x}) \cdot \Pr\{y \geq t_0 | v=0, \mathbf{x}\} \cdot \Pr\{v=0|\mathbf{x}\} \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \Bigg) \\
s.t. & \int_\Omega \left(\Pr\{y \geq t_1 | v=1, \mathbf{x}\} \cdot \Pr\{v=1|\mathbf{x}\} + \Pr\{y \geq t_0 | v=0, \mathbf{x}\} \cdot \Pr\{v=0|\mathbf{x}\} \right) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \leq s \\
\text{and} & \int_\Omega \Pr\{y \geq t_1 | v=1, \mathbf{x}\} \cdot \Pr\{v=1|\mathbf{x}\} \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \geq \frac{r \cdot s \cdot p_1}{1 - p_1 + p_1 \cdot r},
\end{aligned} \tag{20}$$

where t_0 and t_1 are the group-variant quality thresholds, and p_1 is the fraction of minority candidates according to Γ. Given Equation 20, the focus of our analysis is on the difference between $\pi_{\max}$ and the selection quality achieved by an ML algorithm. A substantially lower selection quality from ML would signal the presence of within-group selection bias, and vice versa.

To determine when FAML selection algorithms incur selection bias, we perform a two-step analysis. First, we study the selection outcome when there is no between-group difference in Γ (i.e., $\delta_\Gamma = 0$). Then, we shift our focus to cases with between-group difference (i.e., $\delta_\Gamma > 0$), 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 Γ with between-group difference $\delta_\Gamma = 0$, 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 π_{SE} matching the ideal value $\pi_{\max}$. That is,*

$$\pi_{\mathrm{SE}} = \max_{L:L\in\mathcal{L}} \int_{\Omega} \mathbb{E}_\Gamma(y|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} = \pi_{\max}, \tag{21}$$

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 Γ 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 $\pi_{\max}$. 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 Γ of predictor vector $\mathbf{x}$, group label v, and quality y, such that the between-group difference $\delta_\Gamma \leq d$, and*

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

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

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

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., $\delta_\Gamma \leq 0.5$, to achieve AIR ≥ 0.8, we have $\pi_{\mathrm{SE}}/\pi_{\max} \leq (0.8+1)/(0.8 \cdot 0.25 + 0.8 + 1) = 0.9$ when $s \to 0$, suggesting a loss of at least 10% on selection quality. When the between-group difference is larger, e.g., $\delta_\Gamma = 1$, there is $\pi_{\mathrm{SE}}/\pi_{\max} \leq 0.69$ when $s \to 0$, indicating a loss of over 30% for selection quality. Further, the theorem also shows that the upper bound on $\pi_{\mathrm{SE}}/\pi_{\max}$ 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 $\lim_{s\to 0} \pi_{\mathrm{SE}}/\pi_{\max}$ 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 = $\mathbf{x}$, 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 Finch et al. (2009), using the empirical evidence summarized in Table 1. With this procedure, a dataset was generated in three steps. First, we drew random samples from a multivariate Gaussian distribution with mean being a zero vector and the correlation matrix as specified in Table 1. Each resulting sample is a 6-dimensional vector consisting of both the predictors and the quality y in the table. Second, we randomly assigned each sample a group label v according to an input parameter specifying the proportion of minority candidates. Finally, we added the group difference by subtracting, for each minority candidate, the standardized mean difference value for each predictor and the quality. Following Finch et al. (2009), we verified that the distribution of the generated data was consistent with the specification in Table 1.

5.2.2 Design

Since this procedure does not limit us to only one dataset for analysis, we adjusted various parameters in generating the dataset to test the robustness of our findings in different practical scenarios. Specifically, we varied a total of four parameters. First, we created two levels for the selection rate: $s = .10$ and .20. Second, we created three levels for the ratio between s_1, the retention rate after screening, and the selection rate s: 1, 2, and 3. Note that the case of $s_1/s = 1$ captures the selection task. Third, we created three levels for the fraction of minority candidates: $p_1 = .20$, .40, and .60. Finally, we created nine levels for the fairness constraint: $\text{AIR} = .10, .20, \ldots, .90$.

Overall, our studies consisted of 162 unique conditions or a 2 (s) × 3 (s_1/s) × 3 (p_1) × 9 (AIR) factorial design. Note that we did not manipulate the number of candidates n for two reasons: First, none of the theorems suggests an important role of n. Second, we tested varying levels of n ($n \geq 100$) but found no qualitative differences in the results. Thus, we set a large $n = 1,000$ for all conditions being examined, to reflect the fact that ML algorithms are often used in larger-scale selection scenarios. In terms of the ML algorithm, since we assume the historic training dataset to be sufficiently large so as to reveal all signals about the underlying distribution Γ, we directly calculated Γ from Table 1 before using it to precisely compute the prediction target in Equations 13 and 16, respectively. This way, we could be assured that any degradation of selection quality is caused by the nature of the task rather than prediction errors generated by ML algorithms.

It is important to note that the ML algorithm only has access to the first four predictor variables in Table 1, i.e., biodata, cognitive ability, conscientiousness, and integrity. When ML is assigned the screening task, the manual interview process makes selections according to the fifth predictor variable, i.e., structured interview. Note that the correlation between structured interview and job performance is only .30, reflecting considerable uncertainty in y even at final selection.

5.2.3 Results

Table 2 depicts the mean quality of candidates selected by the ML algorithm under different settings. We compared the ML selection quality with that of the ideal outcomes – i.e., when candidates with the top expected quality in each group are accepted either for selection or screening – in order to gauge any potential fairness issues that may arise from the use of the ML algorithm.

As can be seen from the table, with the selection task, the loss of selection quality (compared with the ideal outcomes) was pervasive and pronounced across all conditions. The average relative loss[8] was 17.70% across all 54 applicable conditions; and the relative loss exceeded 10% in the majority of them (32 out of

[8] This value is different from what was reported in the last row of Table 2 because, due to space limit, Table 2 only included the cases where AIR = .3, .6, or .9, while the values in the text cover all tested conditions.

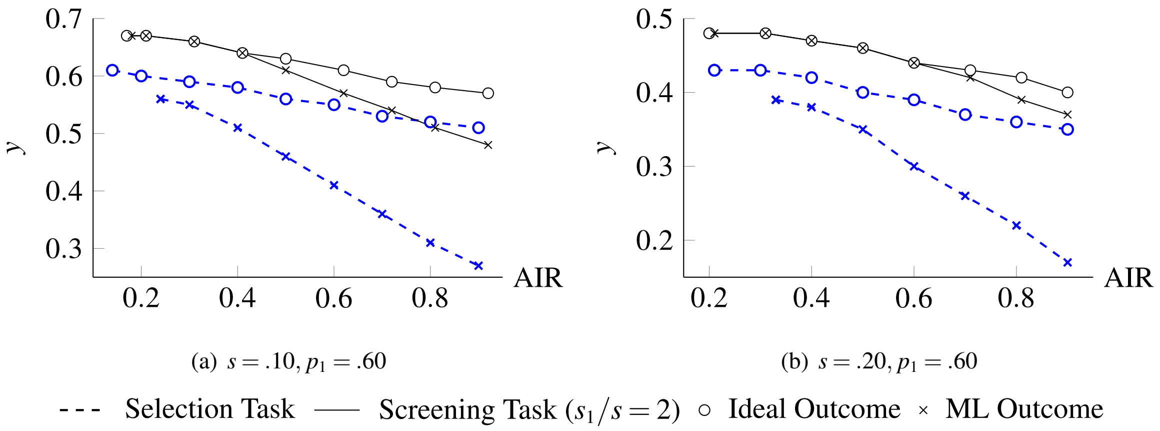
Table 2 Mean Quality of Selected Candidates When AIR = .3, .6, .9

p_1,	$s = .10$											$s = .20$										
	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	δ_1	ID	ML	δ_1	δ_2	ID	ML	δ_1	δ_2	ID	ML	δ_1	ID	ML	δ_1	δ_2	ID	ML	δ_1	δ_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

Note. Standard deviation in parentheses. avg = Average. ID = Ideal outcome, meaning that the decision maker has access to both predictor vector **x** and group label v in selection or screening. ML = outcome using the ML algorithm (which only has access to predictor vector **x**). δ_1 = relative loss of mean selection quality compared with the ideal selection outcome (no screening). δ_2 = relative loss of mean selection quality compared with the selection outcome after ideal screening. Green and red represents cells with relative loss δ_1 smaller than 5% and larger than 20%, respectively. Gray represents those with relative loss in between.

54, 59.26%). This confirms what we proved earlier in the section. That is, instead of accepting the top-quality candidates, the ML algorithm assigned with the selection task chooses candidates with far inferior qualities due to within-group selection bias. For the screening task, however, the relative loss was far lower, averaging only 1.77% across the 108 applicable conditions. Further, the relative loss exceeded 10% for only 2 out of these 108 conditions (1.85%). This confirms that assigning ML with the screening task could effectively address the fairness issues associated with the selection task, shifting the basis of selection back to the quality of selected candidates.

To further illustrate how the selection-screening difference varies with the fairness constraint, we zoom into the worst-case conditions for the ML algorithm in Table 2 (i.e., when $p_1 = 0.6$) and depict in Figure 3 the relationship between selection quality and AIR. As can be seen from the figure, the loss of selection quality is considerably higher for the selection task under a stringent fairness constraint (i.e., a higher AIR). For example, in Figure 3b, with the selection task and an AIR requirement of 0.9, the average selection quality for the ML algorithm is 0.17, over 50% lower than the ideal selection outcome (0.35). In contrast, with the screening task and AIR = 0.9, the ML selection quality is 0.37, less than 10% lower than the ideal outcome (0.40). This confirms our earlier finding that, with the selection task, the ML algorithm tends to



(a) $s = .10, p_1 = .60$

(b) $s = .20, p_1 = .60$

Figure 3 Relationship between Mean Selection Quality and Fairness Constraint

Note. AIR = adverse-impact ratio. The y axis represents the mean quality of all candidates selected. Note that AIR = 0.1 was excluded in certain settings when the selection outcome under AIR ≥ 0.1 yielded AIR > 0.1.

deviate from quality-based selection under stringent fairness constraints, while assigning the ML with the screening task could ameliorate this problem.

5.3 Experimental Studies

We now present experimental studies that examine the last question, i.e., how the FAML selection and screening algorithms compare against each other when both are used in the exact same way to screen candidates for manual interviews. In the passages that follow, we describe the datasets used in the experimental studies, the study design, and the results, respectively.

5.3.1 Datasets

We tested two datasets for the comparison. The first is the simulation dataset described in Section 5.2.1 with $p_1 = 0.2$. The second is a real-world dataset we obtained which contains pre-employment test results, supervisor-rated job performance, and group label (i.e., majority/minority racial group) for 7,890 incumbents of entry-level positions in a Fortune 500 company, including 2,846 (36.07%) protected minorities and 5,044 (63.93%) majorities (as defined by the firm). For each record, there is one quality variable (i.e., y) ranging from 1 to 5, one group indicator (i.e., v) that is either 0 (i.e., majority) or 1 (i.e., minority), and a total of 120 predictor variables (i.e., $\mathbf{x}$) that were collected at the time of hiring. Within the predictor variables, there are 45 that capture the results of situational judgment tests, 20 coded from biodata (i.e., prior experience), and 55 from the results of personality tests. Whereas all variables are integer valued, their scales vary considerably, as some variables represent psychometric assessment scores (e.g., on a 5-point Likert scale) while others represent the number of seconds taken for a candidate to answer a question. We used a random 70%-30% split to form the training and testing dataset, respectively.

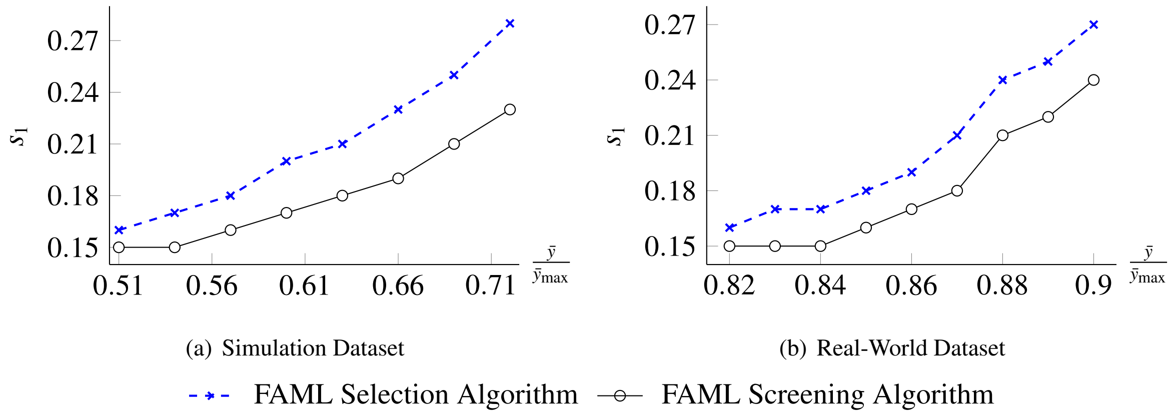
5.3.2 Design of ML Algorithms

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.



(a) Simulation Dataset

(b) Real-World Dataset

Figure 4 Comparison of FAML Selection and Screening Algorithms

Note. s_1 = retention rate specified to the FAML algorithm. $\bar{y}$ = average final selection quality. $\bar{y}_{\max}$ = highest possible value of $\bar{y}$ when $s_1 = 1$. For both panels, the final selection rate was set as $s = 0.1$ with fairness constraint AIR ≥ 1.

Figure 4 shows the results for both datasets. As can be seen from the figure, the screening algorithm outperformed the selection one in all settings. For the simulation dataset, the selection algorithm required a retention rate of 0.28 in order to achieve a final selection quality of $0.72 \cdot \bar{y}_{\max}$, while the screening algorithm only required a retention rate of 0.23, representing a saving of 18% in interview cost. Across all settings, the saving in interview cost ranged from 6% (when $\bar{y}/\bar{y}_{\max} = 0.51$) to 18% (when $\bar{y}/\bar{y}_{\max} = 0.72$). For the real-world dataset, the selection algorithm requires a retention rate of 0.27 in order to achieve a final selection quality of $0.90 \cdot \bar{y}_{\max}$, while the screening algorithm only requires a retention rate of 0.24, representing a saving of 11% in interview cost. Across all settings, the saving in interview cost ranged from 6% (when $\bar{y}/\bar{y}_{\max} = 0.82$) to 14% (when $\bar{y}/\bar{y}_{\max} = 0.87$). In sum, for both datasets, assigning FAML with the screening task (rather than mis-categorizing it as selection) leads to a considerably smaller interview cost for achieving the same level of final selection quality.

6 Discussions

6.1 Policy, Managerial, and Research Implications

In today's "Artificial Intelligence Revolution" (Fuller and Swiontkowski 2020), the growing adoption of ML in organizational decision-making processes makes it increasingly important for researchers, practitioners, and policymakers to be mindful of the interplay between the technical capabilities of ML and the practical constraints occasioned by the legal structure. As discussed earlier, facing the within-group selection bias of FAML selection algorithms, many ML researchers and law scholars have called for a change of law to legalize subgroup norming (e.g., Lipton et al. 2018, Cowgill and Tucker 2020, Wang et al. 2021, Dwork et al. 2018, Bent 2020). Our results in the paper, however, demonstrated that there may be other solutions to the within-group selection bias of existing FAML selection algorithms – e.g., a correct specification of its prediction target according to the screening task rather than oversimplifying the task as selection. To this end, our findings speak to the importance for policymakers *not* to regard the current ML algorithms as finalized products in need of regulatory oversight, but to allow further improvements and refinements through ongoing research.

In terms of managerial implications, our findings suggest that, before using an FAML algorithm in a selection setting in practice, an organization should customize the algorithm according the specific usage scenario. As illustrated in the paper, if reasonable efforts of manual assessments could reveal useful quality signals (even low-validity ones) for selection post ML-screening, an organization could drastically improve the quality of the final selection outcome. As the validity of available predictors could differ substantially across industry and organizational conditions (Song et al. 2017, Kim and Ployhart 2018), our findings suggest that firms should carefully examine the potential predictors and their acquisition costs before designing a proper pipeline that connects ML screening with manual assessment before making the final selection decisions. Our findings also demonstrate that sometimes seemingly trivial changes in ML design – e.g., the simple adjustment of prediction target from Equation 13 to 16 – could lead to substantial improvement in real-world selection scenarios.

For researchers, our work identify new research directions for the use of ML in selection. What we illustrate in the paper is the importance of one operational decision, i.e., whether to assign an ML algorithm with the selection or screening task. In addition to this decision, De Corte et al. (2011) outlined six other design decisions for the operation of a selection system, such as the sequencing of predictors across selection stages (e.g., which to use in screening and which in post-screening selection) and the selection of predictors to be administered based on a given cost constraint. Future studies could investigate how these operational decisions could affect the selection quality and fairness of an FAML algorithm. More broadly, our findings point to the importance of contextualizing the future development of FAML algorithms in realistic selection settings, which could set the stage for more interdisciplinary inquiries into FAML in future research.

6.2 Limitations

Even though we followed the prevailing practice in personnel selection research (e.g., Aguinis et al. 2010, De Corte et al. 2011, Song et al. 2017, Finch et al. 2009) in designing our simulation study, the value of its results is limited by the validity of the empirical evidences reported in the existing meta-analyses, some of which have been challenged in the literature (Morgeson et al. 2007). To this end, we note that the mathematical analysis in the paper (i.e., Theorems 1 and 2) do not make any assumption of the data distribution. Nonetheless, even these mathematical results assume the training-data input to the machine learning algorithm as the ground truth, without taking into account limitations on the training data, such as measurement issues (Hough and Oswald 2000), observational bias (e.g., skewed quality distribution; Lemaître et al. 2017), etc. The prediction errors generated by ML algorithms (cf. epistemic uncertainty for machine learning algorithms; Kendall and Gal 2017) could also affect the validity of our findings. Future 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.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797-806.

Cowgill, Bo, Catherine E Tucker. 2020. Algorithmic fairness and economics. *The Journal of Economic Perspectives*, Forthcoming.

Dastin, Jeffrey. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. San Fransico, CA: Reuters. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/idUSKCN1MK08G`. Accessed: 2022-10-07.

Datta, Anupam, Matthew Fredrikson, Gihyuk Ko, Piotr Mardziel, Shayak Sen. 2017. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. *Proceedings of the 2017 ACM SIGSAC Conference* *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.

Kesselheim, Thomas, Alexandros Psomas, Shai Vardi. 2023. On hiring secretaries with stochastic departures. *Operations Research (Ahead of Print)*, .

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 30 656-666.

Kim, Youngsang, Robert E Ployhart. 2018. The strategic value of selection practices: antecedents and consequences of firm-level selection practice usage. *Academy of Management Journal*, 61 (1), 46-66.

Kleinberg, Jon, Sendhil Mullainathan, Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*. 43:1-43:23.

Kroll, Joshua A, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, Harlan Yu. 2016. Accountable algorithms. *University of Pennsylvania Law Review*, 165 (3), 633-705.

Kusner, Matt J, Joshua Loftus, Chris Russell, Ricardo Silva. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30 4066-4076.

Lemaître, Guillaume, Fernando Nogueira, Christos K Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18 (1), 559-563.

Liem, Cynthia, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, Cornelius J König. 2018. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and interpretable models in computer vision and machine learning*. Springer, 197-253.

Lindley, Denis V. 1961. Dynamic programming and decision theory. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 10 (1), 39-51.

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.

National Research Council. 2004. *Measuring racial discrimination*. National Academies Press.

Newman, David T, Nathanael J Fast, Derek J Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160 149-167.

Nocedal, Jorge, Stephen Wright. 2006. *Numerical optimization*. Springer Science & Business Media.

Oh, Sechan, Özalp Özer. 2016. Characterizing the structure of optimal stopping policies. *Production and Operations Management*, 25 (11), 1820-1838.

Oswald, Frederick L, Eric M Dunleavy, Amy Shaw. 2016. Measuring practical significance in adverse impact analysis. Scott B Morris, Eric M Dunleavy, eds., *Adverse Impact Analysis: Understanding Data, Statistics, and Risk*, 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.

Stewart, Theodor J. 1981. The secretary problem with an unknown number of options. *Operations Research*, 29 (1), 130-145.

Sunar, Nur, Jayashankar M Swaminathan. 2022. Socially relevant and inclusive operations management. *Production and Operations Management*, in press.

Tamaki, Mitsushi. 1991. A secretary problem with uncertain employment and best choice of available candidates. *Operations Research*, 39 (2), 274-284.

Tan, Zilong, Samuel Yeom, Matt Fredrikson, Ameet Talwalkar. 2020. Learning fair representations for kernel models. *International Conference on Artificial Intelligence and Statistics*. PMLR, 155-166.

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: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20 (75), 1-42.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, Cynthia Dwork. 2013. Learning fair representations. *Proceedings of Machine Learning Research*, 28 (3), 325-333.

Zhang, Junzhe, Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. *Advances in Neural Information Processing Systems*, 31 3671-3681.

Zhang, Nan, Mo Wang, Heng Xu, Nick Koenig, Louis Hickman, Jason Kuruzovich, Vincent Ng, Kofi Arhin, Danielle Wilson, Q Chelsea Song, Chen Tang, Leo Alexander III, Yesuel Kim. 2023. Reducing subgroup differences in personnel selection through the application of machine learning. *Personnel Psychology (In Press)*, .

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 race or gender. Due to prevailing laws in the United States (i.e., Civil Rights Act 1991) and the European Union (i.e., Race Equality Directive 2000/43/EC), ML algorithms are *not* allowed to directly access the race or gender of an applicant (so as to avoid the so-called "disparate treatement"; Gottfredson 1994). Yet, as ML researchers soon realized, many other predictors (e.g., test scores for cognitive ability; Hunter and Hunter 1984) contain information about the protected demographic variables, which could be learned by ML algorithms to generate unfair outcomes. As a case in point, Amazon developed an ML algorithm for screening resumes, only to scrap it altogether because the algorithm learned rules (from historic data) that automatically penalize candidates who graduated from women's colleges (Dastin 2018). Similar biases were found for ML algorithms in other selection settings, such as loan underwriting (Fu et al. 2021, Martinez and Kirchner 2021), tenant screening (Gikay 2020), college admission (Burke 2020), etc. To address this crucial concern in designing ML algorithms, researchers started pursuing certain *fairness goals* over the selection outcome generated by the ML algorithm. For example, a common fairness goal for personnel selection is a lower bound on the adverse-impact ratio (AIR), which is the ratio between selection rates for minority and majority candidates (Zhang et al. 2023). ML algorithms that pursue at least one fairness goal are collectively known as *Fairness-Aware ML* (FAML) algorithms, which have seen important contributions from multiple related fields such as operations management (De-Arteaga et al. 2022, Kallus et al. 2022), information systems (Fu et al. 2022), and computer science (Mehrabi et al. 2021).

Nearly all existing FAML algorithms assume that the selection outcome is directly derived from the prediction model generated by the ML algorithm (cf. Mehrabi et al. 2021). In other words, when a firm employs the prediction model to calculate the predicted scores for all applications, it proceeds to select (i.e., approve) those applications with predicted scores above a certain threshold[1]. In practice, however, ML predictions are seldom used as-is in selection settings. Take personnel selection, for example, where ML typically serves a supportive role to Human Resource (HR) managers by screening candidates for subsequent interviews (Liem et al. 2018). That is, ML predictions are primarily utilized to identify and screen out unqualified cases, whereas for the remaining applicants, HR managers manually gather additional information (e.g., through interviews and reference letters) before making decisions. Consequently, the task assigned to ML is more appropriately categorized as a *screening task* rather than a selection task in practice.

Despite this distinction, the prevailing view today is that it has little impact on the design and functioning of FAML algorithms, except for the obvious (quantitative) difference in selection/retention rate. The prevalence of this view is evidenced by the near-universal adoption of the selection task in defining optimization goals for ML algorithms (Mehrabi et al. 2021). On the surface, this view also seems reasonable, as the goal of ML is to assign a "better" application with a higher predicted score. Whether one application is deemed "better" than another seems irrelevant to whether the task at hand is selection or screening.

[1] The threshold could be pre-determined or calculated based on the top-k scores (e.g., if the number of approvals is limited by k).

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] – e.g., either very good or very bad. Opting for such an applicant is inherently risky for the selection task, given the prospect of hiring an unqualified employee. For the screening task, however, the dynamics shift. Even in scenarios where the applicant turns out to be unqualified, a capable HR manager still has the chance to catch and reject the application after an interview, thereby limiting the cost of this risky choice to a wasted interview spot. In other words, an FAML algorithm should logically lean towards making high-risk high-reward choices if it were equipped with the knowledge of (the existence of) latter-stage interviews. Conversely, to design an FAML algorithm for selection, only to subsequently employ it for screening, misses this opportunity afforded by the screening task – an opportunity to leverage the availability of latter-stage interviews to manage the consequences of bold choices. Further, we demonstrate that the ability to make high-risk high-reward choices is crucial for ameliorating the fairness problems known to exist for FAML algorithms, such as selection biases within the minority group and the resulting between-group differences in selection quality (Zhang et al. 2023). This highlights the importance of designing FAML algorithms for the screening task rather than mis-categorizing a screening task as a selection one in algorithmic design.

The rest of the paper is organized as follows. We briefly review the related literature in Section 2. Then, we start by comparing ML selection and screening algorithms without fairness constraint in Section 3, followed by a comparison with fairness constraint in Section 4. Section 5 presents mathematical analysis and experimental results (over both simulated and real-world data) on the comparison between selection and screening. We conclude the paper with discussions of its managerial implications and limitations in Section 6.

2 Literature Review

Organizational selection decisions, such as personnel selection, have been extensively studied across scientific disciplines such as operations management (e.g., Aksin et al. 2007), human resource management (SIOP 2018), machine learning (e.g., Mitchell et al. 2018), etc. In the operations management literature, for example, personnel selection serves as the foundation of the renowned *secretary problem*, which has continued to garner attention over the past half-century (e.g., Lindley 1961, Stewart 1981, Tamaki 1991, Eriksson et al. 2007, Oh and Özer 2016, Kesselheim et al. 2023). More broadly, operations management scholars have delved into various aspects of personnel selection, exploring the link between hiring decisions and employee learning (Arlotto et al. 2014), investigating the complex interactions among hiring, training,

[2] As explained later in the paper, the optimization goals pursued by FAML algorithms make such bimodal distributions common occurrences for FAML predictions.

and turnovers (Aksin et al. 2007), and examining the interface between operations management and human resource management (Boudreau et al. 2003). As our paper focuses on the use of FAML algorithms in selection settings, this literature review zeros in on two issues specific to the use of FAML: 1) what "fairness" means in organizational selection settings; and 2) the existing design of FAML algorithms.

2.1 Fairness Requirements

In terms of what "fairness" means in organizational settings such as hiring and promotions, the prevailing view in both research and practice is that "fairness has no single meaning and, therefore, no single definition" (SIOP 2018). As one cannot exhaust all ideological definitions of fairness, a focus of the existing FAML literature is on satisfying the *legal* mandates that apply to real-world selection settings, especially in the context of the US legal system. In the passages that follow, we review the two main legal mandates, the ban of *disparate treatment* and *disparate impact*, respectively. Note that these legal mandates require a firm to eliminate *both* disparate treatment and disparate impact when making selection decisions.

2.1.1 Disparate Treatment

Disparate-treatment laws prohibit the use of legally protected variables, such as race, ethnicity, gender, national origin, etc., in making selection decisions (Gottfredson 1994, Primus 2010). For example, the US Civil Rights Act (CRA) of 1968 prohibits the use of such protected variables in making lending decisions. Similarly, both the US CRA of 1991 and the European Union Race Equality Directive 2000/43/EC stipulate the same ban in the employment context. The US CRA 1991 explicitly outlaws a then-existing practice of subgroup-norming, making it illegal to "alter the results of employment related tests on the basis of race, color, religion, sex, or national origin" (US CRA 1991, §106).

Enforcing disparate-treatment laws is clearly challenging because many protected variables are readily accessible in selection settings. For example, in job interviews, the protected variables of an interviewee are often directly observable by interviewers, potentially triggering conscious or unconscious biases in their decision making (Purkiss et al. 2006). To prevent such disparate treatment from happening in algorithmic selections, a general consensus in the FAML literature is that FAML algorithms should *not* be allowed to access any protected variable of an applicant (Kroll et al. 2016). In other words, the algorithm should *not* be given information about whether an applicant is in the minority or majority group. We follow this rule throughout the paper.

It is important to understand that banning algorithmic access to protected variables is a necessary but *insufficient* condition for making "fair" selection decisions. A key reason here is that other predictor variables, from location (e.g., ZIP code) to test scores, could well contain information that serves as proxies for protected variables (Pedreshi et al. 2008). In the aforementioned Amazon example, even though the algorithm did not overtly access the gender of any applicant, it did have access to applicants' education background (e.g., attendance in women's colleges), which became a proxy for the protected gender variable. Disparate-impact laws, which we review next, were developed to prevent such "covert" disparity manifested by facially neutral practices.

2.1.2 Disparate Impact

Unlike disparate-treatment laws which regulate the input to an organizational decision process, disparate-impact laws regulate the output of it. Specifically, disparate-impact laws stipulate that there should *not* be a gross statistical disparity in the selection outcome for minority and majority candidates *unless* such disparate impact can be demonstrably justified by a significant, legitimate, business necessity (National Research Council 2004). Further, in the US legal system, a firm may be held liable if there exists an alternative selection process that results in less statistical disparity while serving the firm's legitimate business needs (42 U.S.C. §2000e–2). This creates a clear incentive for firms to reduce any statistical disparity in a selection process (Oswald et al. 2016).

In terms of how to measure the statistical disparity, one of the most widely used metric in practice is the *adverse-impact ratio* (AIR) proposed by the US Equal Employment Opportunity Commission (EEOC) in the *Uniform Guidelines on Employee Selection Procedures* (29 C.F.R. §1607, 1978). Specifically, AIR is defined as the ratio between the selection rate for minority candidates and that for the majorities. Given the wide adoption of AIR in research and practice (De Corte et al. 2011), we use it as the measure of disparate impact in this paper. The larger the AIR, the less disparate impact there is in the selection outcome. Without introducing ambiguity, we use "fairness constraint" to refer to the requirement of AIR $\geq r$, where $r \in [0,1]$ is a pre-determined threshold, throughout the paper.

2.2 FAML Algorithms

In general, an FAML algorithm uses a training dataset – typically consisting of data from past selection decisions – to generate a prediction model for a certain *prediction target* specified to the FAML algorithm. As elaborated in latter sections, the focus of our paper is on how the prediction target should be specified for an FAML algorithm, and whether the specification should differ for selection and screening tasks. The technical design of ML algorithms used to (train a prediction model to) approximate the prediction target, including the functional form of the prediction model, is an issue orthogonal to the focus of this paper. Thus, we keep the review of FAML algorithms brief in the rest of this section.

The FAML literature includes a wide variety of designs, from linear models (Berk et al. 2018) to neural networks (Louizos et al. 2016), from Gaussian processes (Tan et al. 2020) to support-vector machines (Zafar et al. 2019). Many of these designs are mathematically guaranteed to asymptotically converge to the Bayes error when the size of the input training dataset increases (e.g., Gaussian process; Rasmussen and Williams 2006). Since our focus is on the distinction between selection and screening tasks in the prediction target, 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 constraint. We first analyze the optimal design of ML algorithm for selection and screening, respectively. Then, we present an illustrative example to demonstrate the difference in outcome between the two algorithms when both are used to retain the same number of candidates for manual interviews.

3.1 ML for Selection Task

Population of Candidates: Consider a selection setting in which each candidate (e.g., a loan application, a job candidate) is described by $\langle \mathbf{x}, v \rangle$, where $\mathbf{x} \in \Omega$ is a vector of characteristics that can be observed in the selection process (with Ω representing the domain of $\mathbf{x}$), and $v \in \{0, 1\}$ indicates whether the candidate belongs to a protected group (e.g., of racial/ethnic minorities). We refer to individuals with $v = 1$ as the protected *minorities*, and those with $v = 0$ as the majority group. Each candidate is also associated with a non-negative real-valued variable $y \in [0, \infty)$, which represents the candidate's quality of interest. In personnel selection, for example, $\mathbf{x}$ would contain characteristics revealed by a job application, such as an individual's cognitive ability, personality, biodata (i.e., past experience), etc. y would represent the future job 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 Γ 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* as a function of $\mathbf{x}$, namely $L(\mathbf{x}) \in [0,1]$, which describes the probability for a candidate with characteristics $\mathbf{x}$ to be selected. Note that this notation captures the more general setting of stochastic selection decisions. In the special case where ML makes deterministic decisions for a given $\mathbf{x}$, $L(\mathbf{x})$ would simply be limited to one of the two extreme values 0 or 1.

In real-world selection scenarios such as personnel selection, there is usually a pre-determined limit on the fraction of candidates that can be admitted. To capture this limit, we assume the selection decisions to be capacity-constrained, meaning that

$$\int_{\Omega} L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s, \tag{1}$$

where $p_{\Gamma}(\mathbf{x})$ is the marginal probability density function of $\mathbf{x}$ according to the joint distribution Γ, and $s \in [0,1]$ is the *selection rate*, i.e., the maximum fraction of candidates that can be selected.

Subject to the capacity constraint, the ML algorithm is designed to maximize the expected[3] quality of selected candidates. This is equivalent with finding an optimal selection decision function L^* that satisfies

$$\begin{aligned} L^* = \arg\max_{L} \int_{\Omega} \mathbb{E}_{\Gamma}(y|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \\ s.t. \int_{\Omega} L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s, \end{aligned} \tag{2}$$

where $\mathbb{E}_{\Gamma}(y|\mathbf{x})$ is the conditional expectation of y given $\mathbf{x}$ according to the joint distribution Γ.

Optimal ML Design: Rambachan et al. (2020, Proposition 1) proved[4] that the optimal solution to Equation 2 has a strikingly simple form. That is, for any given Γ, there always exists a constant $c \geq 0$, such that $L^*(\mathbf{x}) = 1$ if $\mathbb{E}_{\Gamma}(y|\mathbf{x}) \geq c$ and $L^*(\mathbf{x}) = 0$ otherwise[5]. In other words, if the ML algorithm can accurately estimate the expected quality $\mathbb{E}_{\Gamma}(y|\mathbf{x})$ of each candidate, then its optimal choice would be to simply admit candidates in a decreasing order of their expected quality (i.e., from high to low) until reaching the capacity constraint. Whereas the mathematical proof is subtle, the conclusion is surprisingly consistent with what has been pursued in the field of personnel selection for decades, i.e., to maximize the expected quality (also known as *expected criterion*) of candidates being selected (De Corte et al. 2011).

[3] Note that expectation is taken over the randomness of both the latent quality variable y and the stochastic selection decisions $L(\cdot)$.

[4] At a high level, the proof can be considered an extension of the classic Hardy-Littlewood inequality (Hardy et al. 1952).

[5] This statement assumes a finite density function for the marginal distribution of y according to Γ. If the assumption is violated, then we might need a tie-breaking design of $L^*(\mathbf{x}) \in [0,1]$ when $\mathbb{E}_{\Gamma}(y|\mathbf{x}) = c$ to satisfy the capacity constraint. See Rambachan et al.'s (2020) proof for details.

With this, the ML algorithm design is then reduced to generating an accurate point estimate of $\mathbb{E}_\Gamma(y|\mathbf{x})$ for a given $\mathbf{x}$. To do so, the ML algorithm learns from a training dataset formed by historic instances of $\langle \mathbf{x}, v, y\rangle$ which are assumed to be drawn from the same joint distribution Γ. 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 $\langle \mathbf{x}, v, y\rangle$ to be sufficiently large so as to allow ML to learn the joint distribution Γ 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 $\langle \mathbf{x}, v, y\rangle$ drawn from joint distribution Γ, the ML algorithm only has access to the candidate's characteristics vector $\mathbf{x}$, and admits the candidate with probability $L(\mathbf{x}) \in [0,1]$. Unlike in the selection setting, the candidates admitted by the ML algorithm are not directly selected. Instead, they enter a second-stage interview process in which the final selection decisions are made by human experts (e.g., HR managers). This change has two main implications in the design of ML algorithm.

First, an analysis of the final selection quality now requires a model of the manual interview process. In practice, human experts may gather additional information beyond the characteristics in $\mathbf{x}$ during the interview process. They may also opt to collect different information for different candidates. Such flexibility makes the *process* of manual interview extremely difficult to model. To address the challenge, we instead model the *outcome* of the interview process denoted by a binary variable $\mathbb{T} \in \{0,1\}$, with $\mathbb{T} = 1$ representing a candidate passing the interview and $\mathbb{T} = 0$ otherwise.

Since the purpose of this paper is to examine the design of ML algorithms for screening, we consider manual interviews that yield the best possible selection outcome given the ML screening results. In other words, the manual interview selects an optimal subset (in terms of expected quality) of candidates who passed ML screening (subject to the selection rate constraint). Obviously, the optimal subset is formed by candidates with quality y over a certain cutoff y_0. That is, $\mathbb{T} = \mathbb{1}(y \geq y_0)$, where $\mathbb{1}(y \geq y_0)$ is the indicator function that returns 1 if $y \geq y_0$ and 0 otherwise. With this, the pursuit of selection quality is equivalent with maximizing

$$u = \int_\Omega \mathbb{E}_\Gamma(y \cdot \mathbb{T}|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \tag{3}$$

$$= \int_{\Omega} \mathbb{E}_{\Gamma}(y \cdot \mathbb{1}(y \geq y_0)|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \tag{4}$$

$$= \int_{\Omega} \mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x}) \cdot \Pr\{y \geq y_0|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x}, \tag{5}$$

where $\mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x})$ is the expectation of y given $\mathbf{x}$ conditional upon $y \geq y_0$. Intuitively, Equation 5 suggests that, in the screening case, only those candidates who can pass the manual interview matters for the final selection quality.

Second, the introduction of manual interviews also alters the capacity constraints. There are now two such constraints governing ML screening and final selection, respectively. For ML screening, there must be

$$\int_{\Omega} L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s_1, \tag{6}$$

where s_1 is the maximum fraction of candidates that can be retained to manual interviews. Then, the final selection must obey

$$\int_{\Omega} \Pr\{y \geq y_0|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s, \tag{7}$$

where s is the final selection rate.

Putting together Equation 5 with the two capacity constraints, we see that the objective of the ML algorithm under the screening setting is to find

$$L^* = \arg\max_{L} \int_{\Omega} \mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x}) \cdot \Pr\{y \geq y_0|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x}$$

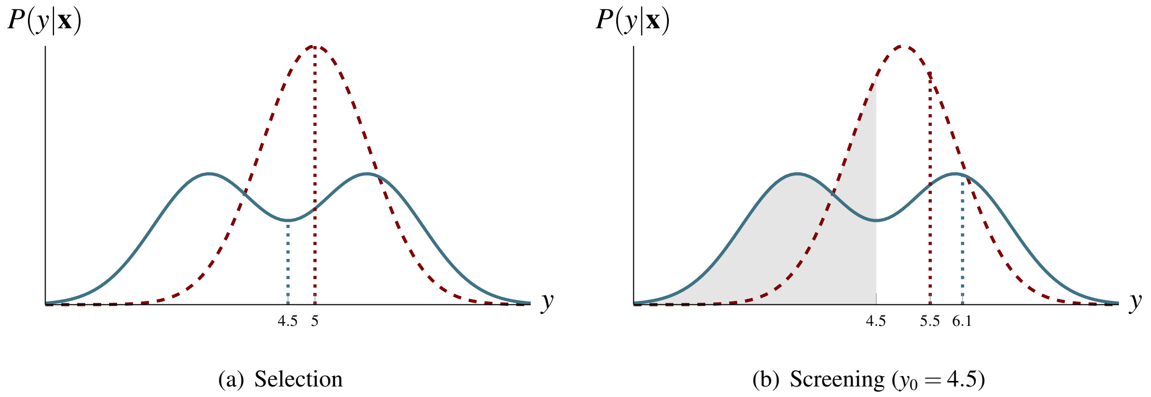
$$s.t. \int_{\Omega} L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s_1 \text{ and } \int_{\Omega} \Pr\{y \geq y_0|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_{\Gamma}(\mathbf{x}) d\mathbf{x} \leq s. \tag{8}$$

Optimal ML Design: With two capacity constraints, the optimization problem becomes considerably more difficult to solve compared with the selection case. To ease the discussions, we start with a simplifying assumption that the interview cost is low – i.e., s_1 is sufficiently large so as to make Inequality 7 the only capacity constraint that matters – before removing this assumption later in mathematical and experimental analyses. Note that this simplification does not imply a trivial ML solution of retaining all candidates (i.e., setting $L^*(\mathbf{x}) = 1$ for all $\mathbf{x}$) because doing so may violate the capacity constraint on final selections (i.e., Inequality 7). With the simplification, Rambachan et al.'s (2020) proof directly carries over to the screening case, with the only change (from the selection case) being the replacement of $\mathbb{E}_{\Gamma}(y|\mathbf{x})$ with $\mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x})$. In other words, under the screening setting, the optimal choice for the ML algorithm is to retain candidates with characteristics $\mathbf{x}$ in a decreasing order of $\mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x})$ – i.e., their expected quality *conditional upon passing the manual interview* – until reaching the capacity constraint.

Compared with the selection case, the difference is straightforward. In the selection case, the expected quality of every ML-admitted candidate matters (hence the ranking of $\mathbb{E}_{\Gamma}(y|\mathbf{x})$), because they all affect the final selection quality. In the screening case, however, only those ML-admitted candidates who can pass the manual interview matters (hence the ranking of $\mathbb{E}_{\Gamma}(y|y \geq y_0, \mathbf{x})$), because the others affect only interview costs but not the final selection quality. As elaborated next, this key difference gives a screening-oriented ML algorithm the ability to make high-risk high-reward choices in retaining candidates.

3.3 Comparison between Selection and Screening

Whereas the previous two subsections explicate the design difference of ML algorithms for selection and screening, we now examine how such design differences lead to different outcomes when both algorithms are used in the exact 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. For each candidate, both algorithms have access to the exact same information, i.e., characteristics $\mathbf{x}$ and nothing else. Both algorithms are also given the same training dataset formed by historic instances of $\langle \mathbf{x}, v, y \rangle$. The selection algorithm is simply specified to choose s_1 fraction of candidates. The screening algorithm is specified to do the same, but is also given s as input, with the understanding that s out of the s_1 retained candidates will be eventually selected. This way, the screening algorithm can compute the quality cutoff y_0 accordingly.



(a) Selection

(b) Screening ($y_0 = 4.5$)

Figure 1 Illustrative Example for Selection vs. Screening without Fairness Constraint

Note. Both panels depict the probability density function of $P(y|\mathbf{x})$ for Alice (red dashed) and Bob (blue solid). Panel (a) shows that Alice has a higher expected quality (i.e., $\mathbb{E}_\Gamma(y|\mathbf{x}) = 5$) than Bob (4.5), and is thus the preferred choice in the selection setting. For the screening setting, Panel (b) shows, when we consider the expected quality conditional upon passing a subsequent interview with $y_0 = 4.5$, then Bob becomes the preferred choice with $\mathbb{E}_\Gamma(y|y \geq y_0, \mathbf{x}) = 6.1$ over Alice (5.5). The shaded region in Panel (b) represents scenarios where Alice or Bob is filtered out by the manual interview.

To highlight differences between the two algorithms, we consider an illustrative example comparing two candidates Alice and Bob, and explain why the two algorithms would have different preferences between them. Recall from earlier discussions that both algorithms make their decisions about a candidate based entirely on the conditional probability distribution of quality y given the candidate's characteristics $\mathbf{x}$ (i.e., $P(y|\mathbf{x})$), which we refer to as the *quality distribution* of a candidate. In the example, we consider quality distributions for Alice and Bob as depicted in Figure 1. For Alice, the distribution is Gaussian $N(5,1)$, i.e., with mean 5 and variance 1. For Bob, the distribution is a Gaussian mixture with two components[6] of equal weight, one being $N(3,1)$ and the other being $N(6,1)$. As can be seen from the figure, Bob's quality

[6] As elaborated in the next section, such bimodal Gaussian mixture is indeed common occurrence in quality distribution for FAML.

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 $\mathbb{E}_\Gamma(y|\mathbf{x})$ 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., $\mathbb{E}_\Gamma(y|y \geq y_0, \mathbf{x})$), 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 be filtered out by the interview anyway, affecting only the interview cost yet having zero effect on the final selection quality. This de-risking feature of manual interview is what allows the screening algorithm to make high-risk high-reward choices (like Bob) that the selection algorithm is unable to make.

In sum, even when both selection and screening algorithms are used in the same way (i.e., to retain s_1 fraction of candidates), they could reach different conclusions on whether one candidate is "better" than another, and therefore produce different outcomes. As elaborated in the next section, this difference is a key contributor to the known fairness issues incurred by FAML algorithms designed for the selection setting.

4 Fairness Implications of Selection vs. Screening

We now examine the differences between ML for selection and screening *with* the presence of fairness constraint. Like in the last section, we first analyze the optimal design of FAML algorithms for selection and screening, respectively, before using an illustrative example to explain their difference when both are used to retain candidates for manual interviews. Further, we show that this difference directly addresses the known fairness issues incurred by FAML algorithms designed for the selection setting.

4.1 Selection with Fairness Constraint

Recall from Section 2.1.2 that we focus on fairness constraint expressed as a lower bound on the adverse impact ratio (AIR), i.e., AIR $\geq r$, where AIR is the ratio between the selection rates of minorities and majorities. For example, a constraint of AIR ≥ 1 would require the selection rate for minorities to be at least as high as that for majorities. Given the fraction of minority candidates p_1 (according to the joint distribution Γ) and the selection rate s, simple algebraic transformations can reduce AIR $\geq r$ to

$$\int_\Omega \Pr\{v = 1|\mathbf{x}\} \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x})d\mathbf{x} \geq \frac{r \cdot s \cdot p_1}{1 - p_1 + p_1 \cdot r}, \tag{9}$$

where $\Pr\{v=1|\mathbf{x}\}$ is the conditional probability of a candidate being a minority given $\mathbf{x}$ according to Γ, because otherwise there would be

$$\text{AIR} = \frac{\frac{1}{p_1}\cdot\int_\Omega \Pr\{v=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x}}{\frac{1}{1-p_1}\cdot\left(s-\int_\Omega \Pr\{v=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x}\right)} < \frac{\frac{r\cdot s}{1-p_1+p_1\cdot r}}{\frac{1}{1-p_1}\cdot\frac{s-s\cdot p_1}{1-p_1+p_1\cdot r}} = r. \tag{10}$$

Putting together this new AIR constraint with Equation 2, the objective of an FAML algorithm (i.e., with fairness constraint) becomes to find

$$\begin{aligned} &L^* = \arg\max_L \int_\Omega \mathbb{E}_\Gamma(y|\mathbf{x})\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \\ &s.t. \int_\Omega L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s \text{ and } \int_\Omega \Pr\{v=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \ge \frac{r\cdot s\cdot p_1}{1-p_1+p_1\cdot r} \end{aligned} \tag{11}$$

Zhang et al. (2023, Theorem 1) proved that the optimal solution to Equation 11 can be deduced through the method of Lagrange multiplier (Nocedal and Wright 2006). That is, for any given s, p_1, and r, there always exists a Lagrange multiplier $\lambda \ge 0$, such that Equation 11 is equivalent with

$$\begin{aligned} &L^* = \arg\max_L \int_\Omega \left(\mathbb{E}_\Gamma(y|\mathbf{x}) + \lambda\cdot\Pr\{v=1|\mathbf{x}\}\right)\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \\ &s.t. \int_\Omega L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s. \end{aligned} \tag{12}$$

Clearly, the larger r is, the larger λ will be. When $r=0$, we have $\lambda=0$, reducing Equation 12 to the baseline selection case without fairness constraint.

With this transformation, Rambachan et al.'s (2020) Proposition 1 can again be directly applied. That is, under the selection setting with fairness constraint, the optimal choice for the FAML algorithm is to admit candidates with characteristics $\mathbf{x}$ in a decreasing order of

$$f(\mathbf{x}) = \mathbb{E}_\Gamma(y|\mathbf{x}) + \lambda\cdot\Pr\{v=1|\mathbf{x}\} \tag{13}$$

until reaching the capacity constraint. Compared with the selection case without fairness constraint, Equation 13 includes a new additive component of $\lambda\cdot\Pr\{v=1|\mathbf{x}\}$. In other words, instead of predicting quality alone, the FAML algorithm is designed to predict a weighted sum of the expected quality and the probability for the candidate to be a minority.

4.2 Screening with Fairness Constraint

A similar transformation can be carried out for the screening task with fairness constraint. Specifically, putting together the selection quality in Equation 3 with the fairness constraint AIR $\ge r$, we see that the objective of an FAML screening algorithm is to find

$$\begin{aligned} &L^* = \arg\max_L \int_\Omega \mathbb{E}_\Gamma(y\cdot\mathbb{T}|\mathbf{x})\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \\ &s.t. \int_\Omega L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s_1, \int_\Omega \Pr\{\mathbb{T}=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s, \\ &\text{and } \int_\Omega \Pr\{v=1,\mathbb{T}=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \ge \frac{r\cdot s\cdot p_1}{1-p_1+p_1\cdot r}, \end{aligned} \tag{14}$$

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

$$\begin{aligned}
L^* &= \arg\max_L \int_\Omega (\mathbb{E}_\Gamma(y\cdot\mathbb{T}|\mathbf{x}) + \lambda\cdot\Pr\{v=1,\mathbb{T}=1|\mathbf{x}\})\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \\
&= \arg\max_L \int_\Omega (\mathbb{E}_\Gamma(y|\mathbb{T}=1,\mathbf{x})\cdot\Pr\{\mathbb{T}=1|\mathbf{x}\} + \lambda\cdot\Pr\{v=1|\mathbb{T}=1,\mathbf{x}\}\cdot\Pr\{\mathbb{T}=1|\mathbf{x}\})\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \\
&= \arg\max_L \int_\Omega (\mathbb{E}_\Gamma(y|\mathbb{T}=1,\mathbf{x}) + \lambda\cdot\Pr\{v=1|\mathbb{T}=1,\mathbf{x}\})\cdot\Pr\{\mathbb{T}=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \\
s.t.& \int_\Omega \Pr\{\mathbb{T}=1|\mathbf{x}\}\cdot L(\mathbf{x})\cdot p_\Gamma(\mathbf{x})d\mathbf{x} \le s, \qquad (15)
\end{aligned}$$

where λ ($\lambda \ge 0$) 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

$$f'(\mathbf{x}) = \mathbb{E}_\Gamma(y|\mathbb{T}=1,\mathbf{x}) + \lambda\cdot\Pr\{v=1|\mathbb{T}=1,\mathbf{x}\} \qquad (16)$$

until reaching the capacity constraint.

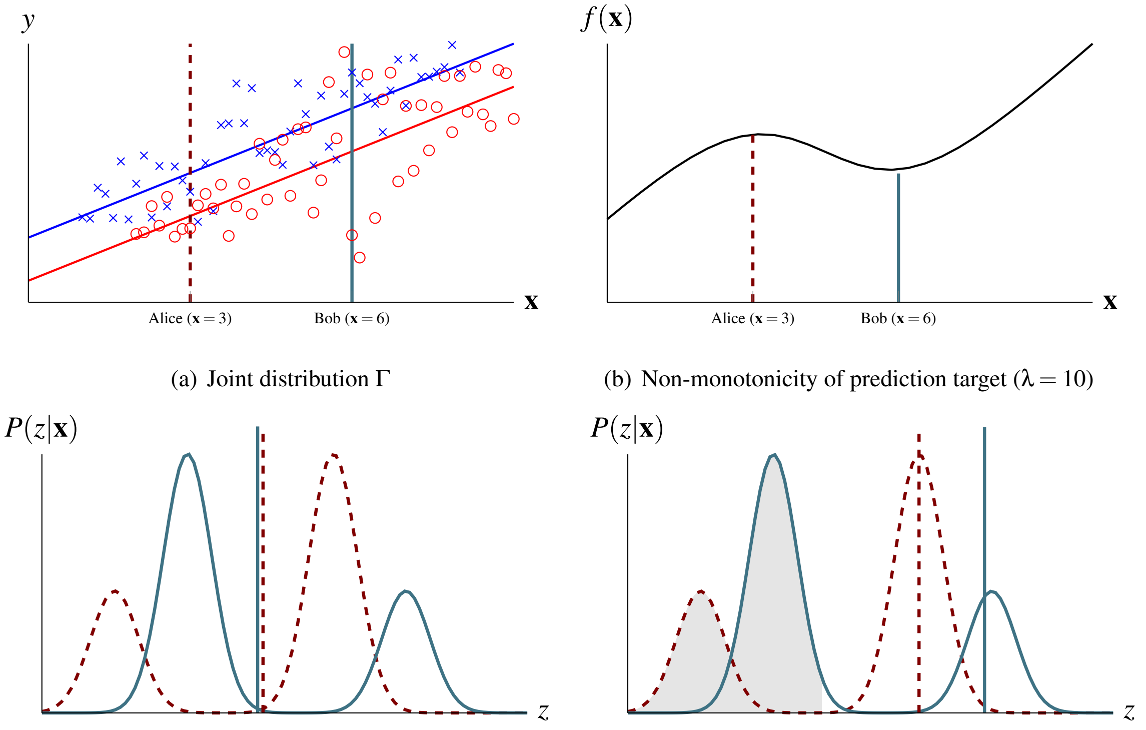
Juxtaposing Equation 16 with the optimal design for the selection case (i.e., Equation 13), the difference is, in essence, the same as the selection-screening difference without fairness constraint. That is, for screening, only candidates who can pass the manual interview matters for final selection 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 $\mathbb{T} = \mathbb{1}(y+\lambda_2 v \ge t_0)$, 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

$$f'(\mathbf{x}) = \mathbb{E}_\Gamma(y|y+\lambda_2 v \ge t_0,\mathbf{x}) + \lambda\cdot\Pr\{v=1|y+\lambda_2 v\ge t_0,\mathbf{x}\} \qquad (17)$$

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 ($s \le s_1$) fraction of candidates who must satisfy the fairness constraint of AIR $\ge r$. 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



(a) Joint distribution Γ

(b) Non-monotonicity of prediction target ($\lambda = 10$)

(c) Selection ($z = y + \lambda \cdot v$, $\lambda = 10$)

(d) Screening ($\lambda = \lambda_2 = 10, t_0 = 8$)

Figure 2 Illustrative Example for Selection vs. Screening with Fairness Constraint.

Note. Panel (a) depicts the joint distribution Γ, with $\times$ and ○ representing minorities and majorities, respectively. Panel (b) shows that $f(\mathbf{x})$, the prediction target for FAML selection algorithm, does not monotically increase with $\mathbf{x}$. Panels (c) and (d) depict the probability density function of $P(z|\mathbf{x})$ for Alice (red dashed) and Bob (blue solid), where $z = y + \lambda \cdot v$. The two vertical lines in Panel (c) show the expected value of z (i.e., the prediction target for selection algorithm) for Alice and Bob, while those in Panel (d) show the conditional expectation of z given $z \geq t_0$ (i.e., the prediction target for screening algorithm). The shaded region in Panel (d) represents scenarios where Alice or Bob is filtered out by the manual interview (i.e., $z < t_0$).

interview, it has to be specified to choose s_1 fraction of candidates while satisfying the fairness constraint (i.e., AIR $\geq r$) for the chosen candidates (as is done in almost all existing work; Mehrabi et al. 2021). The screening algorithm, thanks to it accounting for the manual interview, is specified to choose s_1 fraction of candidates, but is also given the final selection rate s as input with the requirement that the final selection outcome satisfying AIR $\geq r$.

To highlight differences between the two algorithms, we once again consider their preferences between two candidates, Alice and Bob. To make the comparison more realistic, we no longer use hypothetical quality distributions as in the earlier example (Section 3.3). Instead, we start with defining a simple joint distribution Γ according to empirical findings in the personnel selection literature. Specifically, ample empirical evidence suggests that, compared with a majority candidate with the same quality, a minority candidate tends to score lower in predictors that are positively correlated with quality (e.g., SAT scores for college admission, cognitive ability tests in personnel selection; Berry et al. 2011), with a possible reason being that 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,

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

where $\varepsilon \sim N(0,1)$ 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 *de-risks* the selection of Bob because his left (i.e., "risky") component is now mostly filtered out by the manual interview. As a result, the screening algorithm no longer needs to skip a higher-quality candidate (i.e., Bob) to choose a low-risk alternative (i.e., Alice). Indicatively, the screening prediction target for Bob (14.70) is now larger than Alice (12.00), meaning that the FAML algorithm for screening prefers Bob over Alice, consistent with the order of their expected quality. As can be seen from this example, it is the screening algorithm's ability to make high-risk high-reward choices that ameliorates the within-group selection bias of FAML selection algorithms. In other words, it is crucial to properly specify the screening task to an FAML algorithm rather than miscategorizing it as a selection task.

5 Mathematical Analysis, Simulation, and Experimental Studies

Conceptual development in the previous two sections reveals a difference between FAML selection and screening algorithms in their risk-taking tendency and, consequently, their preferences between different candidates. This finding raises three questions to be answered through mathematical analysis and experimental studies. The first question is on whether within-group selection bias (depicted in Figure 2b) always manifests as quality degradation when FAML algorithms are directly used in a selection setting. Answering this question could help determine whether it is necessary to incur the cost of manual interviews in practice. We use mathematical analysis to study this question in the first part of this section.

Then, the second question is on whether assigning FAML with the screening task followed by manual interviews could ameliorate the quality degradation caused by within-group selection bias. Answering this question could help us understand whether addressing the within-group selection bias of FAML selection algorithms requires a change of the current legal system (as argued in existing work, e.g., Lipton et al. 2018), or if such bias could be alleviated by simply correcting the mis-categorization of screening task as a selection one. We examine this second question through simulation studies in the second part of this section.

Finally, once we establish the superiority of using FAML followed by manual interviews, the third question that arises is how the actual performance of FAML selection and screening algorithms differ when *both* are used to screen candidates for manual interviews. Since this comparison depends on not only the data distribution but the ML algorithm being used, we address it through experimental studies on both simulated and real-world data in the last part of this section.

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 Γ. Specifically, we are interested in between-group difference on $P(y|\mathbf{x})$, the conditional distribution of y given **x**, because FAML selection algorithm relies on $P(y|\mathbf{x})$ in their decision-making. To capture between-group difference on $P(y|\mathbf{x})$, 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 Γ is defined as

$$\delta_\Gamma = \max_{\Theta \subseteq \Omega} \left| \frac{\mathbb{E}_\Gamma(y|\mathbf{x} \in \Theta, v=0) - \mathbb{E}_\Gamma(y|\mathbf{x} \in \Theta, v=1)}{\mathrm{SD}_\Gamma(y|\mathbf{x} \in \Theta)} \right|, \tag{19}$$

where Ω is the domain of **x**, $|\cdot|$ is the absolute value, and SD_Γ represents standard deviation over Γ.

In terms of the value of δ_Γ in practice, Roth et al. (2003) show that between-group difference varies depending on the quality measure being used, e.g., from 0.13 for a subjective measure of absenteeism to 0.52 for an objective measure of work samples to 0.55 for an objective measure of job knowledge.

Besides δ_Γ, we also need a way to detect the within-group selection bias discussed in Section 4.3. Recall that such a bias manifests as a reduction of selection quality because, within the minority (or majority) group, it could favor a candidate with lower expected quality over a higher-quality candidate. Thus, we detect the presence of within-group selection bias by comparing the final selection quality of FAML algorithms with the maximum possible selection quality subject to capacity (i.e., selection rate s) and fairness (i.e., AIR $\geq r$) constraints. As discussed in Section 4.2, this ideal selection quality can be captured by

$$\begin{aligned}
\pi_{\max} = \max_{t_0, t_1} \Bigg(&\int_\Omega \mathbb{E}_\Gamma(y|y \geq t_1, v=1, \mathbf{x}) \cdot \Pr\{y \geq t_1 | v=1, \mathbf{x}\} \cdot \Pr\{v=1|\mathbf{x}\} \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} + \\
&\int_\Omega \mathbb{E}_\Gamma(y|y \geq t_0, v=0, \mathbf{x}) \cdot \Pr\{y \geq t_0 | v=0, \mathbf{x}\} \cdot \Pr\{v=0|\mathbf{x}\} \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \Bigg) \\
s.t. &\int_\Omega \left(\Pr\{y \geq t_1 | v=1, \mathbf{x}\} \cdot \Pr\{v=1|\mathbf{x}\} + \Pr\{y \geq t_0 | v=0, \mathbf{x}\} \cdot \Pr\{v=0|\mathbf{x}\}\right) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \leq s \\
\text{and} &\int_\Omega \Pr\{y \geq t_1 | v=1, \mathbf{x}\} \cdot \Pr\{v=1|\mathbf{x}\} \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} \geq \frac{r \cdot s \cdot p_1}{1 - p_1 + p_1 \cdot r},
\end{aligned} \tag{20}$$

where t_0 and t_1 are the group-variant quality thresholds, and p_1 is the fraction of minority candidates according to Γ. Given Equation 20, the focus of our analysis is on the difference between $\pi_{\max}$ and the selection quality achieved by an ML algorithm. A substantially lower selection quality from ML would signal the presence of within-group selection bias, and vice versa.

To determine when FAML selection algorithms incur selection bias, we perform a two-step analysis. First, we study the selection outcome when there is no between-group difference in Γ (i.e., $\delta_\Gamma = 0$). Then, we shift our focus to cases with between-group difference (i.e., $\delta_\Gamma > 0$), 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 Γ with between-group difference $\delta_\Gamma = 0$, 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 π_{SE} matching the ideal value $\pi_{\max}$. That is,*

$$\pi_{\text{SE}} = \max_{L:L\in\mathcal{L}} \int_{\Omega} \mathbb{E}_\Gamma(y|\mathbf{x}) \cdot L(\mathbf{x}) \cdot p_\Gamma(\mathbf{x}) d\mathbf{x} = \pi_{\max}, \tag{21}$$

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 Γ 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 $\pi_{\max}$. 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 Γ of predictor vector $\mathbf{x}$, group label v, and quality y, such that the between-group difference $\delta_\Gamma \leq d$, and*

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

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

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

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., $\delta_\Gamma \leq 0.5$, to achieve AIR ≥ 0.8, we have $\pi_{\text{SE}}/\pi_{\max} \leq (0.8+1)/(0.8 \cdot 0.25 + 0.8 + 1) = 0.9$ when $s \to 0$, suggesting a loss of at least 10% on selection quality. When the between-group difference is larger, e.g., $\delta_\Gamma = 1$, there is $\pi_{\text{SE}}/\pi_{\max} \leq 0.69$ when $s \to 0$, indicating a loss of over 30% for selection quality. Further, the theorem also shows that the upper bound on $\pi_{\text{SE}}/\pi_{\max}$ 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 $\lim_{s\to 0} \pi_{\text{SE}}/\pi_{\max}$ 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 = $\mathbf{x}$, 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 Finch et al. (2009), using the empirical evidence summarized in Table 1. With this procedure, a dataset was generated in three steps. First, we drew random samples from a multivariate Gaussian distribution with mean being a zero vector and the correlation matrix as specified in Table 1. Each resulting sample is a 6-dimensional vector consisting of both the predictors and the quality y in the table. Second, we randomly assigned each sample a group label v according to an input parameter specifying the proportion of minority candidates. Finally, we added the group difference by subtracting, for each minority candidate, the standardized mean difference value for each predictor and the quality. Following Finch et al. (2009), we verified that the distribution of the generated data was consistent with the specification in Table 1.

5.2.2 Design

Since this procedure does not limit us to only one dataset for analysis, we adjusted various parameters in generating the dataset to test the robustness of our findings in different practical scenarios. Specifically, we varied a total of four parameters. First, we created two levels for the selection rate: $s = .10$ and .20. Second, we created three levels for the ratio between s_1, the retention rate after screening, and the selection rate s: 1, 2, and 3. Note that the case of $s_1/s = 1$ captures the selection task. Third, we created three levels for the fraction of minority candidates: $p_1 = .20$, .40, and .60. Finally, we created nine levels for the fairness constraint: $\text{AIR} = .10, .20, \ldots, .90$.

Overall, our studies consisted of 162 unique conditions or a 2 (s) × 3 (s_1/s) × 3 (p_1) × 9 (AIR) factorial design. Note that we did not manipulate the number of candidates n for two reasons: First, none of the theorems suggests an important role of n. Second, we tested varying levels of n ($n \geq 100$) but found no qualitative differences in the results. Thus, we set a large $n = 1,000$ for all conditions being examined, to reflect the fact that ML algorithms are often used in larger-scale selection scenarios. In terms of the ML algorithm, since we assume the historic training dataset to be sufficiently large so as to reveal all signals about the underlying distribution Γ, we directly calculated Γ from Table 1 before using it to precisely compute the prediction target in Equations 13 and 16, respectively. This way, we could be assured that any degradation of selection quality is caused by the nature of the task rather than prediction errors generated by ML algorithms.

It is important to note that the ML algorithm only has access to the first four predictor variables in Table 1, i.e., biodata, cognitive ability, conscientiousness, and integrity. When ML is assigned the screening task, the manual interview process makes selections according to the fifth predictor variable, i.e., structured interview. Note that the correlation between structured interview and job performance is only .30, reflecting considerable uncertainty in y even at final selection.

5.2.3 Results

Table 2 depicts the mean quality of candidates selected by the ML algorithm under different settings. We compared the ML selection quality with that of the ideal outcomes – i.e., when candidates with the top expected quality in each group are accepted either for selection or screening – in order to gauge any potential fairness issues that may arise from the use of the ML algorithm.

As can be seen from the table, with the selection task, the loss of selection quality (compared with the ideal outcomes) was pervasive and pronounced across all conditions. The average relative loss[8] was 17.70% across all 54 applicable conditions; and the relative loss exceeded 10% in the majority of them (32 out of

[8] This value is different from what was reported in the last row of Table 2 because, due to space limit, Table 2 only included the cases where AIR = .3, .6, or .9, while the values in the text cover all tested conditions.

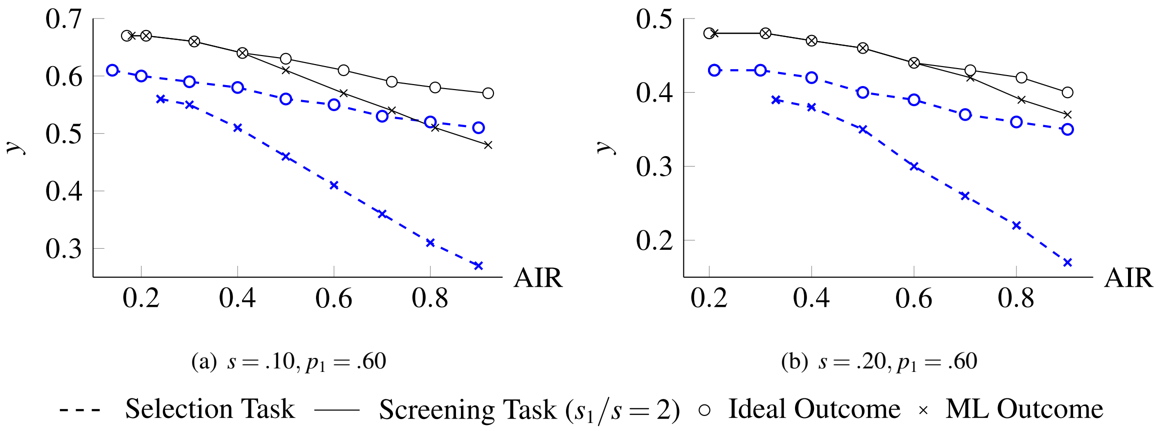
Table 2 Mean Quality of Selected Candidates When AIR = .3, .6, .9

	$s = .10$											$s = .20$										
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	δ_1	ID	ML	δ_1	δ_2	ID	ML	δ_1	δ_2	ID	ML	δ_1	ID	ML	δ_1	δ_2	ID	ML	δ_1	δ_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

Note. Standard deviation in parentheses. avg = Average. ID = Ideal outcome, meaning that the decision maker has access to both predictor vector **x** and group label v in selection or screening. ML = outcome using the ML algorithm (which only has access to predictor vector **x**). δ_1 = relative loss of mean selection quality compared with the ideal selection outcome (no screening). δ_2 = relative loss of mean selection quality compared with the selection outcome after ideal screening. Green and red represents cells with relative loss δ_1 smaller than 5% and larger than 20%, respectively. Gray represents those with relative loss in between.

54, 59.26%). This confirms what we proved earlier in the section. That is, instead of accepting the top-quality candidates, the ML algorithm assigned with the selection task chooses candidates with far inferior qualities due to within-group selection bias. For the screening task, however, the relative loss was far lower, averaging only 1.77% across the 108 applicable conditions. Further, the relative loss exceeded 10% for only 2 out of these 108 conditions (1.85%). This confirms that assigning ML with the screening task could effectively address the fairness issues associated with the selection task, shifting the basis of selection back to the quality of selected candidates.

To further illustrate how the selection-screening difference varies with the fairness constraint, we zoom into the worst-case conditions for the ML algorithm in Table 2 (i.e., when $p_1 = 0.6$) and depict in Figure 3 the relationship between selection quality and AIR. As can be seen from the figure, the loss of selection quality is considerably higher for the selection task under a stringent fairness constraint (i.e., a higher AIR). For example, in Figure 3b, with the selection task and an AIR requirement of 0.9, the average selection quality for the ML algorithm is 0.17, over 50% lower than the ideal selection outcome (0.35). In contrast, with the screening task and AIR = 0.9, the ML selection quality is 0.37, less than 10% lower than the ideal outcome (0.40). This confirms our earlier finding that, with the selection task, the ML algorithm tends to



(a) $s = .10, p_1 = .60$

(b) $s = .20, p_1 = .60$

Figure 3 Relationship between Mean Selection Quality and Fairness Constraint

Note. AIR = adverse-impact ratio. The y axis represents the mean quality of all candidates selected. Note that AIR = 0.1 was excluded in certain settings when the selection outcome under AIR ≥ 0.1 yielded AIR > 0.1.

deviate from quality-based selection under stringent fairness constraints, while assigning the ML with the screening task could ameliorate this problem.

5.3 Experimental Studies

We now present experimental studies that examine the last question, i.e., how the FAML selection and screening algorithms compare against each other when both are used in the exact same way to screen candidates for manual interviews. In the passages that follow, we describe the datasets used in the experimental studies, the study design, and the results, respectively.

5.3.1 Datasets

We tested two datasets for the comparison. The first is the simulation dataset described in Section 5.2.1 with $p_1 = 0.2$. The second is a real-world dataset we obtained which contains pre-employment test results, supervisor-rated job performance, and group label (i.e., majority/minority racial group) for 7,890 incumbents of entry-level positions in a Fortune 500 company, including 2,846 (36.07%) protected minorities and 5,044 (63.93%) majorities (as defined by the firm). For each record, there is one quality variable (i.e., y) ranging from 1 to 5, one group indicator (i.e., v) that is either 0 (i.e., majority) or 1 (i.e., minority), and a total of 120 predictor variables (i.e., $\mathbf{x}$) that were collected at the time of hiring. Within the predictor variables, there are 45 that capture the results of situational judgment tests, 20 coded from biodata (i.e., prior experience), and 55 from the results of personality tests. Whereas all variables are integer valued, their scales vary considerably, as some variables represent psychometric assessment scores (e.g., on a 5-point Likert scale) while others represent the number of seconds taken for a candidate to answer a question. We used a random 70%-30% split to form the training and testing dataset, respectively.

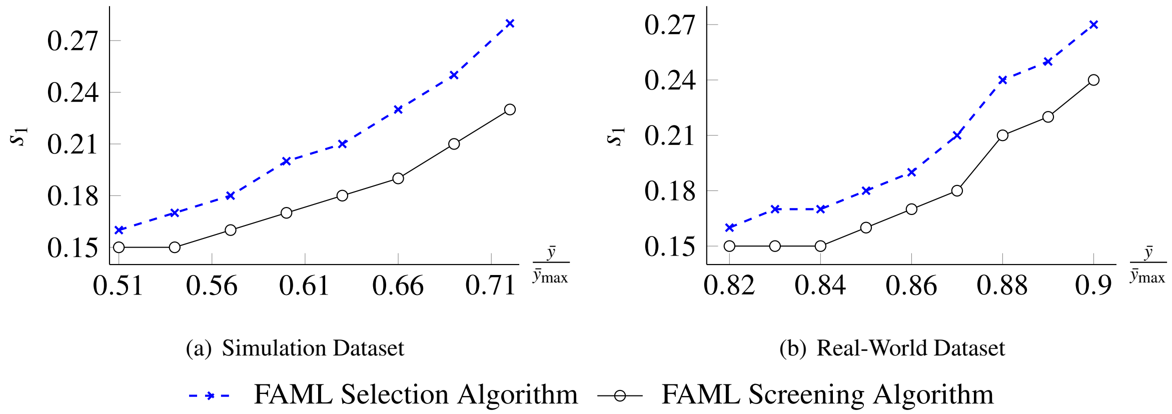
5.3.2 Design of ML Algorithms

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.



(a) Simulation Dataset

(b) Real-World Dataset

Figure 4 Comparison of FAML Selection and Screening Algorithms

Note. s_1 = retention rate specified to the FAML algorithm. $\bar{y}$ = average final selection quality. $\bar{y}_{\max}$ = highest possible value of $\bar{y}$ when $s_1 = 1$. For both panels, the final selection rate was set as $s = 0.1$ with fairness constraint AIR ≥ 1.

Figure 4 shows the results for both datasets. As can be seen from the figure, the screening algorithm outperformed the selection one in all settings. For the simulation dataset, the selection algorithm required a retention rate of 0.28 in order to achieve a final selection quality of $0.72 \cdot \bar{y}_{\max}$, while the screening algorithm only required a retention rate of 0.23, representing a saving of 18% in interview cost. Across all settings, the saving in interview cost ranged from 6% (when $\bar{y}/\bar{y}_{\max} = 0.51$) to 18% (when $\bar{y}/\bar{y}_{\max} = 0.72$). For the real-world dataset, the selection algorithm requires a retention rate of 0.27 in order to achieve a final selection quality of $0.90 \cdot \bar{y}_{\max}$, while the screening algorithm only requires a retention rate of 0.24, representing a saving of 11% in interview cost. Across all settings, the saving in interview cost ranged from 6% (when $\bar{y}/\bar{y}_{\max} = 0.82$) to 14% (when $\bar{y}/\bar{y}_{\max} = 0.87$). In sum, for both datasets, assigning FAML with the screening task (rather than mis-categorizing it as selection) leads to a considerably smaller interview cost for achieving the same level of final selection quality.

6 Discussions

6.1 Policy, Managerial, and Research Implications

In today's "Artificial Intelligence Revolution" (Fuller and Swiontkowski 2020), the growing adoption of ML in organizational decision-making processes makes it increasingly important for researchers, practitioners, and policymakers to be mindful of the interplay between the technical capabilities of ML and the practical constraints occasioned by the legal structure. As discussed earlier, facing the within-group selection bias of FAML selection algorithms, many ML researchers and law scholars have called for a change of law to legalize subgroup norming (e.g., Lipton et al. 2018, Cowgill and Tucker 2020, Wang et al. 2021, Dwork et al. 2018, Bent 2020). Our results in the paper, however, demonstrated that there may be other solutions to the within-group selection bias of existing FAML selection algorithms – e.g., a correct specification of its prediction target according to the screening task rather than oversimplifying the task as selection. To this end, our findings speak to the importance for policymakers *not* to regard the current ML algorithms as finalized products in need of regulatory oversight, but to allow further improvements and refinements through ongoing research.

In terms of managerial implications, our findings suggest that, before using an FAML algorithm in a selection setting in practice, an organization should customize the algorithm according the specific usage scenario. As illustrated in the paper, if reasonable efforts of manual assessments could reveal useful quality signals (even low-validity ones) for selection post ML-screening, an organization could drastically improve the quality of the final selection outcome. As the validity of available predictors could differ substantially across industry and organizational conditions (Song et al. 2017, Kim and Ployhart 2018), our findings suggest that firms should carefully examine the potential predictors and their acquisition costs before designing a proper pipeline that connects ML screening with manual assessment before making the final selection decisions. Our findings also demonstrate that sometimes seemingly trivial changes in ML design – e.g., the simple adjustment of prediction target from Equation 13 to 16 – could lead to substantial improvement in real-world selection scenarios.

For researchers, our work identify new research directions for the use of ML in selection. What we illustrate in the paper is the importance of one operational decision, i.e., whether to assign an ML algorithm with the selection or screening task. In addition to this decision, De Corte et al. (2011) outlined six other design decisions for the operation of a selection system, such as the sequencing of predictors across selection stages (e.g., which to use in screening and which in post-screening selection) and the selection of predictors to be administered based on a given cost constraint. Future studies could investigate how these operational decisions could affect the selection quality and fairness of an FAML algorithm. More broadly, our findings point to the importance of contextualizing the future development of FAML algorithms in realistic selection settings, which could set the stage for more interdisciplinary inquiries into FAML in future research.

6.2 Limitations

Even though we followed the prevailing practice in personnel selection research (e.g., Aguinis et al. 2010, De Corte et al. 2011, Song et al. 2017, Finch et al. 2009) in designing our simulation study, the value of its results is limited by the validity of the empirical evidences reported in the existing meta-analyses, some of which have been challenged in the literature (Morgeson et al. 2007). To this end, we note that the mathematical analysis in the paper (i.e., Theorems 1 and 2) do not make any assumption of the data distribution. Nonetheless, even these mathematical results assume the training-data input to the machine learning algorithm as the ground truth, without taking into account limitations on the training data, such as measurement issues (Hough and Oswald 2000), observational bias (e.g., skewed quality distribution; Lemaître et al. 2017), etc. The prediction errors generated by ML algorithms (cf. epistemic uncertainty for machine learning algorithms; Kendall and Gal 2017) could also affect the validity of our findings. Future 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.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797-806.

Cowgill, Bo, Catherine E Tucker. 2020. Algorithmic fairness and economics. *The Journal of Economic Perspectives*, Forthcoming.

Dastin, Jeffrey. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. San Fransico, CA: Reuters. `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/idUSKCN1MK08G`. Accessed: 2022-10-07.

Datta, Anupam, Matthew Fredrikson, Gihyuk Ko, Piotr Mardziel, Shayak Sen. 2017. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. *Proceedings of the 2017 ACM SIGSAC Conference* *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.

Kesselheim, Thomas, Alexandros Psomas, Shai Vardi. 2023. On hiring secretaries with stochastic departures. *Operations Research (Ahead of Print)*, .

Kilbertus, Niki, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 30 656-666.

Kim, Youngsang, Robert E Ployhart. 2018. The strategic value of selection practices: antecedents and consequences of firm-level selection practice usage. *Academy of Management Journal*, 61 (1), 46-66.

Kleinberg, Jon, Sendhil Mullainathan, Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*. 43:1-43:23.

Kroll, Joshua A, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, Harlan Yu. 2016. Accountable algorithms. *University of Pennsylvania Law Review*, 165 (3), 633-705.

Kusner, Matt J, Joshua Loftus, Chris Russell, Ricardo Silva. 2017. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30 4066-4076.

Lemaître, Guillaume, Fernando Nogueira, Christos K Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18 (1), 559-563.

Liem, Cynthia, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, Cornelius J König. 2018. Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. *Explainable and interpretable models in computer vision and machine learning*. Springer, 197-253.

Lindley, Denis V. 1961. Dynamic programming and decision theory. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 10 (1), 39-51.

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.

National Research Council. 2004. *Measuring racial discrimination*. National Academies Press.

Newman, David T, Nathanael J Fast, Derek J Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160 149-167.

Nocedal, Jorge, Stephen Wright. 2006. *Numerical optimization*. Springer Science & Business Media.

Oh, Sechan, Özalp Özer. 2016. Characterizing the structure of optimal stopping policies. *Production and Operations Management*, 25 (11), 1820-1838.

Oswald, Frederick L, Eric M Dunleavy, Amy Shaw. 2016. Measuring practical significance in adverse impact analysis. Scott B Morris, Eric M Dunleavy, eds., *Adverse Impact Analysis: Understanding Data, Statistics, and Risk*, 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.

Stewart, Theodor J. 1981. The secretary problem with an unknown number of options. *Operations Research*, 29 (1), 130-145.

Sunar, Nur, Jayashankar M Swaminathan. 2022. Socially relevant and inclusive operations management. *Production and Operations Management*, in press.

Tamaki, Mitsushi. 1991. A secretary problem with uncertain employment and best choice of available candidates. *Operations Research*, 39 (2), 274-284.

Tan, Zilong, Samuel Yeom, Matt Fredrikson, Ameet Talwalkar. 2020. Learning fair representations for kernel models. *International Conference on Artificial Intelligence and Statistics*. PMLR, 155-166.

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: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20 (75), 1-42.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, Cynthia Dwork. 2013. Learning fair representations. *Proceedings of Machine Learning Research*, 28 (3), 325-333.

Zhang, Junzhe, Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. *Advances in Neural Information Processing Systems*, 31 3671-3681.

Zhang, Nan, Mo Wang, Heng Xu, Nick Koenig, Louis Hickman, Jason Kuruzovich, Vincent Ng, Kofi Arhin, Danielle Wilson, Q Chelsea Song, Chen Tang, Leo Alexander III, Yesuel Kim. 2023. Reducing subgroup differences in personnel selection through the application of machine learning. *Personnel Psychology (In Press)*, .