Hiring with Algorithmic Fairness Constraints: Theory and Empirics

Prasanna Parasurama* Panos Ipeirotis

New York University – Stern School of Business

March 23, 2023

Abstract

In algorithmic hiring systems, diversity policies are often inscribed as algorithmic fairness constraints. But algorithms rarely work in isolation; almost always, humans make the ultimate hiring decision based on recommendations from the algorithm. To better understand the downstream effects of algorithmic fairness constraints in Human+AI hiring systems, we present, solve and empirically estimate a 2-stage hiring model consisting of an algorithmic screener and an unbiased hiring manager. In the first stage, an algorithmic screener screens candidates from a pool of applicants consisting of equally-qualified men and women. The algorithm estimates candidate quality and shortlists the best candidates based on this estimate. There is a parity constraint (i.e., a diversity policy) imposed on the algorithm to shortlist an equal number of men and women. In the second stage, the hiring manager estimates candidate quality again and hires the best candidate based on this estimate from the shortlist. We solve this model analytically and show that even when both the algorithm and the hiring manager are unbiased, the parity constraint can be ineffective in increasing the gender diversity of the hires. The effectiveness of the parity constraint depends on parameters such as the size of the applicant pool, the proportion of women in the applicant pool, gender differences in the screening algorithm's predictive power, and most importantly, the correlation between the algorithm's and the hiring manager's assessment of candidate quality. The *more* correlated the screening algorithm's and the hiring manager's quality estimates are, the *less* effective the parity constraint becomes in increasing workforce diversity. We empirically estimate these parameters using hiring data from IT firms and show via counterfactual policy simulation that parity constraint can improve the *average* proportion of female hires by a modest amount; however, there will be a high level of heterogeneity in the effectiveness across job types. We discuss the managerial and algorithmic design implications of these findings.

*pparasurama@stern.nyu.edu

1 Introduction

In recent years, many firms have adopted various diversity policies to increase the diversity of their workforce (Shi et al. 2018). One such policy is to diversify the shortlist/interview pool, sometimes called a "soft" affirmative action policy[1]. An example is the NFL's Rooney Rule[2], a hiring policy that requires the leagues to interview at least one ethnic minority for the head coach position. Similar policies have since been adopted by various large tech firms, including Patreon[3], Pinterest[4], and Facebook to increase the racial and gender diversity of their workforce. For example, in 2016, Facebook implemented a point system for recruiters to source diverse candidates (Huet 2017). Under this system, Facebook recruiters received 1 point if their candidate got hired and an additional bonus point if the hired candidate was diverse. However, the effectiveness of this policy is in question due to the misalignment of incentives between recruiters and engineers/managers who ultimately made the hiring decision. Managers did not get extra incentives to hire diverse candidates, so many of the diverse candidates that the recruiters brought in did not get hired. As a result, recruiters soon reverted to usual recruitment strategies, ignoring the diversity bonus altogether.

> The recruiters saw that many of their diversity candidates didn't get an offer. Two former recruiters blamed in part the engineering department's candidate review process... Getting "diversity candidates" hired at Facebook proved to be such a struggle that many recruiters stopped trying, even with the double point system, and went back to their usual strategies (Huet 2017).

As hiring becomes increasingly aided by algorithms, codifying such diversity policies in the form of algorithmic fairness constraints may sound promising. Unlike Facebook recruiters, algorithmic screeners are guaranteed not to ignore the diversity policy. At the same time, algorithms rarely work in isolation; almost always, a human decision-maker (such as a hiring manager) makes the ultimate hiring decision based on recommendations from the algorithm. So, the *overall* effectiveness of these

[1] "Soft" affirmative action policies aim to increase the share of minority candidates in the initial stages of the hiring funnel, but with no "hard" quotas on hiring. This stands in contrast with "hard" affirmative action policies, which require explicit and mandatory consideration of minority status in hiring decisions to reach a quota. In most cases, hard affirmative action policies are prohibited under U.S employment law (Civil Rights Act of 1974). See Schuck (2002).

[2] https://operations.nfl.com/inside-football-ops/diversity-inclusion/the-rooney-rule/

[3] https://www.slideshare.net/TarynArnold/patreon-culture-deck-april-2017

[4] https://newsroom.pinterest.com/en/post/our-plan-for-a-more-diverse-pinterest

constraints is far from a guarantee but depends on how algorithmic recommendations interact with human decisions. For example, in 2018, *LinkedIn Recruiter*[5] started using a fairness-aware ranking algorithm to improve the gender diversity of candidates shown to recruiters (Geyik et al. 2019). The authors show that the introduction of fairness-aware ranking did not significantly affect business metrics such as the number of LinkedIn messages sent. However, it is unclear whether improving gender representation in ranking improved *outcomes* (such as more messages, interview requests) for underrepresented candidates (Geyik et al. 2019).

Prior lab studies have shown that the effectiveness of such fairness constraints depends vastly on the job type (Sühr et al. 2021; Peng et al. 2019). When these fairness constraints or diversity policies do not work, the conventional wisdom is to attribute the ineffectiveness to the bias of human decision-makers. Indeed human bias can negate any fairness constraints in an algorithm, but there is little understanding of what *other* factors contribute to the effectiveness of fairness constraints.

To identify these factors in a structured manner, we first develop a 2-stage hiring model consisting of an algorithmic and human component. In the first stage of the model, a screening algorithm screens and shortlists candidates from a pool of applicants. There are more male than female applicants, but both have the same underlying quality distribution. To improve the gender diversity of the workforce, the algorithm has a gender-parity constraint such that it shortlists an equal number of men and women. In the second stage, an *unbiased* hiring manager interviews the shortlisted candidates and hires the best candidate based on her assessment. We solve this model analytically and show that the parity constraint can be ineffective in increasing the gender diversity of hires *even when the hiring manager is unbiased*. We identify three parameters that determine the effectiveness of the parity constraint in balancing the gender proportion of hires: (1) the size of the applicant pool, (2) gender differences in the screening algorithm's predictive power, and (3) the correlation between the algorithm's and the hiring manager's assessment of candidate quality. Interestingly, the effectiveness of the fairness constraint decreases as the correlation between the algorithm's and the hiring manager's estimates of candidate quality increases. The better the screening algorithm learns the hiring manager's assessment criteria, the less effective the fairness constraint becomes.

In the paper's second half, we empirically estimate the above model parameters and test the theoretical predictions on real-world hiring data from eight technology firms (799k applicants, 3.6k

[5]LinkedIn's sourcing and talent search tool for recruiters.

job postings). Specifically, we train two deep-learning models: (1) a resume screening model, which learns from screeners' decisions, and (2) a hiring manager model, which learns from managers' decisions. Using the predictions from these models, we estimate the parameters of our 2-stage hiring model. We then perform counterfactual policy simulation to understand the effectiveness of parity constraint and show that: (1) In line with theoretical predictions, parity constraint in the shortlist does not necessarily lead to gender parity in hires and, in some instances, does not affect the outcome. (2) While parity constraint helps improve the *average* proportion of female hires by a modest amount, there is a high level of heterogeneity in effectiveness across job types.

In addressing these research questions, we seek to make two main contributions to the literature. First, we make a theoretical contribution that furthers our understanding of how various non-bias-related factors contribute to the overall effectiveness of algorithmic fairness constraints when algorithmic outputs are used as inputs in downstream decisions. Second, we make an empirical/methodological contribution in using ML-driven agent-based simulations to achieve more realistic counterfactual policy simulations – a methodological innovation that is beginning to get adopted in economics (see, e.g., Zheng et al. (2020)).

The rest of the paper is organized as follows. Section 2 discusses related works. Section 3 introduces the theoretical hiring model. Section 4 presents the theoretical results and proofs. Section 5 introduces the data and empirical modeling techniques. Section 6 presents the empirical results. Section 7 discusses the managerial implications of the results and concludes.

2 Related Work

Researchers have studied and documented algorithmic bias in various contexts including hiring, lending, and criminal justice (Datta et al. 2015; Lambrecht and Tucker 2019; Dastin 2018; Angwin and Larson 2022; Chouldechova 2017). The algorithmic fairness literature concerns the design of fair algorithms to mitigate such bias (Dwork et al. 2012; Zemel et al. 2013; Hardt et al. 2016; Zafar, Valera, Gomez Rodriguez, et al. 2017; Zafar, Valera, Rodriguez, et al. 2017; Geyik et al. 2019; Blum et al. 2022). In this literature, bias and fairness are generally considered along the lines of *protected attributes*[6] such as race, gender, religion, sexual orientation, etc. Different definitions

[6]Attributes that are typically protected under the law against discrimination. For example, U.S Federal law prohibits employment discrimination based on race, gender, religion, national origin, age, disability, sexual orientation,

of fairness exist, such as equal false-positive rates, equal false-negative rates, equal odds (ratio of false positive rate and false negative rates), equal accuracy rates, and equal positive predictive values across groups (See Mitchell et al. (2021) for a review). Researchers have shown that it is impossible to simultaneously satisfy all fairness criteria except for in trivial cases (Chouldechova 2017; Kleinberg, Mullainathan, et al. 2016), so the choice of fairness criteria depends on the context and is often informed by existing laws and policies[7]. *Demographic parity* is a commonly used fairness criterion in the algorithmic hiring setting, where the proportion of positive outcomes across groups is constrained to be equal to the proportions in a baseline population (Raghavan et al. 2020). For example, in algorithmic screening, demographic parity may require that the proportion of women on the shortlist be equal to the proportion of women in the applicant pool (here, the applicant pool is the baseline population). However, there is no universally agreed-upon definition of the baseline population. In LinkedIn's fairness-aware ranking, Geyik et al. (2019) take the subset of qualified candidates as the baseline population, while Sühr et al. (2021) take the entire labor force of an online labor market (TaskRabbit) as the baseline population.

Fairness constraints are not only used to mitigate any potential bias in the algorithm but are often used as a tool to proactively increase the diversity of candidates. For example, LinkedIn's fairness-aware ranking algorithm improves gender diversity in candidate ranking in *LinkedIn Recruiter* compared to other ranking algorithms that only consider efficiency (Geyik et al. 2019). When algorithmic recommendations are used as inputs to human decisions, the downstream effects of these fairness constraints in either mitigating bias or increasing diversity are not guaranteed, as the net effect depends on how the algorithmic recommendations interact with human decisions. For example, Teodorescu et al. (2021) highlight the challenges of automating algorithmic fairness when humans are involved. Within the hiring context, Sühr et al. (2021) study the effectiveness of fairness constraints on hiring outcomes to understand whether fair ranking improves minority outcomes in a lab setting. They find that while fair ranking improved minority outcomes in some jobs, the effect is dampened in jobs with persistent gender preferences, such as moving assistance jobs. In a similar survey study, Peng et al. (2019) manipulate the gender distribution of candidates shown to participants and find that increasing representation in the candidate pool can correct for biases in

and pregnancy.

[7] In employment, EEOC's 4/5ths rule is used as a guideline, which states that the selection rate for all protected groups must be at least 4/5ths of the group with the highest selection rate.

some jobs. But in heavily gendered jobs such as urologist, OBGYN, and nanny, manipulating the gender representation in the candidate pool has no impact.

Some theoretical papers have also studied the downstream effects of manipulating the gender distribution in the candidate pool on hiring outcomes. Kleinberg and Raghavan (2018) provide a theoretical hiring model in the presence of implicit bias. They show that the Rooney Rule can increase the proportion of minority hires while also increasing the payoff of the decision-maker. Celis et al. (2021) study the effects of the Rooney Rule in the long-term with learning and show that the Rooney Rule reduces implicit bias and the reduction rate increases with the size of the candidate pool relative to the size of the shortlist. Both models assume that the decision-maker is biased. Fershtman and Pavan (2021) present a model to study the effect of "soft" affirmative action policies that increase the proportion of minority candidates in the candidate pool. They show that such policies can have no effect or even backfire if the evaluation of minority candidates is noisier than that of majority candidates, even in the absence of bias. L. M. Lee and Waddell (2021) study a 2-stage hiring setting with agents with different levels of interest in diversity. They show that this difference can lower the likelihood of highly qualified candidates to be hired even when they enhance diversity.

3 Model

Consider a hiring context in which a firm seeks to hire someone for a job. There are n_a applicants for the job, and each applicant belongs to a group $g \in \{m, f\}$, where m is the majority group and f is the minority group. Proportion p of the n_a applicants belong to the minority group ($p < 0.5$), and $1 - p$ belong to the majority group. For exposition, we label the majority group *male* and the minority group *female*.

Each candidate has a true quality Q that is unobserved at hiring but revealed once hired (e.g., future job performance). We model Q as a random variable drawn from an underlying distribution F_Q. We assume that both male and female applicants have the same quality distribution – i.e., there is no expected quality difference between a randomly chosen male and female.

The firm implements a resume screening algorithm that screens and shortlists n_s candidates. To increase the gender diversity of its workforce, the firm imposes a gender parity constraint on the

screening algorithm such that it shortlists an equal number of men and women. For now, assume that the algorithm is statistically unbiased and equally predictive for male and female candidates alike (we will relax this assumption later).

In the first stage, the screening algorithm screens n_a resumes and estimates a quality score Q^S (a noisy signal of the candidate's true quality score Q) based on the characteristics in the resume, rank orders them based on Q^S, and shortlists the top n_s candidates. Under parity constraint, the algorithm will shortlist the top $\frac{n_s}{2}$ male candidates and the top $\frac{n_s}{2}$ female candidates. Note that the scores themselves are independent of gender $Q^S \perp g$, even though the shortlist decisions depend on gender.

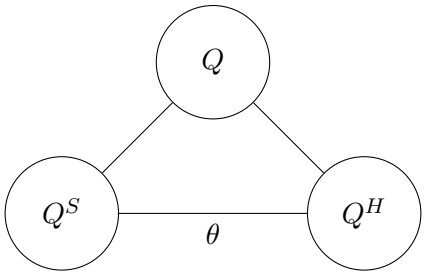
Without loss of generality, we model the screening quality score Q^S as a random variable drawn from a uniform distribution. Regardless of what the true distribution of Q is, it can be mapped to a uniform distribution using the integral transform, such that the best candidate has a score close to 1 and the worst candidate has a score of 0. Because we assume both men and women have the same quality distribution, the distribution of screening scores will also be the same.

$$\begin{aligned} Q^S_{male} &\sim U[0,1] \\ Q^S_{female} &\sim U[0,1] \end{aligned} \tag{3.1}$$

In the second stage, an *unbiased* human hiring manager interviews the shortlisted candidates, comes up with her own estimate of quality Q^H for each candidate after the interview, rank orders them based on Q^H, and hires the best candidate (number of hires $n_h = 1$). Unlike the screening algorithm, the hiring manager has no gender parity constraints.

We allow the screening algorithm's estimate of the quality Q^S and the hiring manager's estimate of the quality Q^H to be correlated with a "correlation" parameter $\theta \in [0,1]$[8]. When $\theta = 1$, the screening and the hiring manager's scores will be perfectly correlated. When $\theta = 0$, the screening score and the hiring manager's will be perfectly orthogonal (we discuss this in the following section).

[8] In a slight abuse of terminology, we call this the correlation parameter for clarity. Dependency parameter would be a more accurate term but is not as clear.



For now, consider the screening algorithm as a black-box model that returns a score Q^S given some job and candidate characteristics, and disregard how the screening algorithm is trained and how Q^S is computed. The following theoretical results are agnostic to how to screening algorithm is trained, but we will discuss this in detail in Section 4.7.

Given this model, our goal is to find the probability of a female candidate getting hired as a function of the model parameters (n_a, n_s, θ, p). To solve for this probability, we will need the distribution of the hiring manager scores of the shortlisted candidates. To derive this distribution, we first need to know the distribution of screening scores of the shortlisted candidate and the dependency structure that relates screening scores to hiring manager scores.

Table 1: Table of Notations

Symbol	Definition
n_a	Number of applicants
n_s	Number of candidates in the shortlist
n_h	Number of hires
p	Proportion of females in the applicant pool ($p < 0.5$)
Q	Candidate's true quality score (Unobserved)
Q^S	Screening score (noisy estimate of true quality)
Q_s^S	Screening score of the shortlisted candidate
$\hat{q}_i^S$	Estimate of screening score for candidate i
Q^H	Hiring manager score (noisy estimate of true quality)
Q_s^H	Hiring manager score of the shortlisted candidate
$\hat{q}_i^H$	Estimate of hiring manager score for candidate i
θ	Correlation between Q^S and Q^H
δ	Gender difference in correlation $\theta_m - \theta_f$
θ^S	Correlation between Q^S and Q
θ^H	Correlation between Q^H and Q

A note on notations: Capital letters denote random variables, small letters denote variables, Greek letters denote parameters to estimate, $\hat{x}$ denotes the estimate of x. Generally, subscripts denote some subset of candidates. For example $Q_{s,m}$ denotes the quality (Q) of the shortlisted (s) male (m) candidates.

3.1 Dependency between Screening and Hiring Manager Scores

The dependency between screening score Q^S and the hiring manager score Q^H is a modeling choice. We consider two different dependency structures: (1) Gaussian Copula and (2) Mixture Distribution. A key desirable property in choosing a dependency structure is that it must allow us to model the dependency between the two scores with a single parameter (which we label θ)[9]. The exact interpretation of θ varies based on the dependency structure, but in both cases, θ ranges from 0 to 1, which is meant to capture the "strength" of the dependency. Our main results are not sensitive to the choice of the dependency structure.

3.1.1 Gaussian Copula

We first consider the copula dependency structure – a popular method to model dependencies between distributions. Copulas are multivariate distributions, whose marginals are uniformly distributed (See Joe (2014) and Nelsen (2007) for a reference on copulas).

Definition (Copula)**.** *Copula $C_{U,V}(u, v)$ is the joint distribution of U and V, where U and V are uniform random variables.*

Even if the marginals are not uniformly distributed, they can be transformed into a uniform distribution using the integral transform (i.e., taking the inverse CDF of any distribution). Therefore, copulas offer a flexible way to disentangle any joint distribution as a product of univariate marginal distributions and a copula that "couples" them. They are often used in cases where the joint distribution is unknown or unimportant. In our case, we use copulas because they allow us to model the dependency between the two scores with a single parameter θ.

To see why copulas can be useful, consider a random vector (X, Y) whose marginal distributions are $F_X(x)$, $F_Y(y)$. We know there is a dependency between X and Y, but we do not know the joint distribution $F_{X,Y}(x, y)$. Instead, we model the dependency with a copula using Sklar's theorem, which states that any joint distribution can be rewritten using a copula.

Theorem (Sklar's Theorem)**.** *For any joint distribution $F_{X,Y}(x, y)$, there exists a copula C such that $F_{X,Y}(x, y) = C(F_X(x), F_Y(y))$*

[9] Although joint distributions may be a natural way to model dependencies, they don't allow us to model the dependency with a single parameter.

Given one of the marginals and the copula, we can recover the other marginal as such:

$$f_Y(y) = \int_X c(x, y) \cdot f_X(x) dx$$

where $c(x, y)$ is density function of the copula. Applying this to the model at hand, given the marginal distribution of the screening score F_{Q^S} and a copula that models the dependency, we can recover the marginal distribution of the hiring manager's score F_{Q^H}.

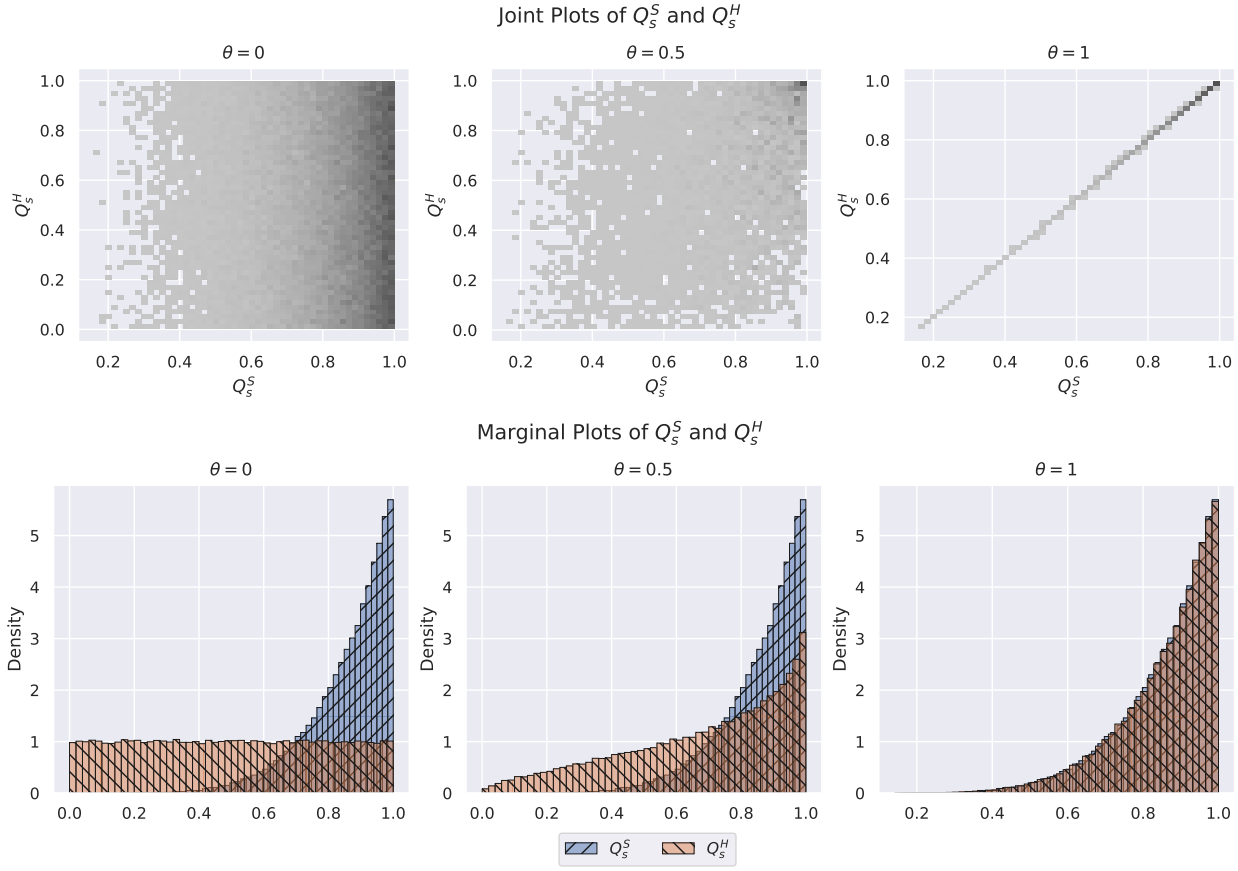
$$f_{Q^H}(y) = \int_{Q^S} c(x, y) \cdot f_{Q^S}(x) dx \tag{3.2}$$

Given that copulas are multivariate distributions, there are infinitely many, some of which are useful. The Gaussian copula is a commonly used copula, which, as the name suggests, is a multivariate Gaussian distribution.

Definition (Gaussian Copula)**.** *C is a Gaussian copula parameterized by θ, where $C(u, v; \theta) = \Phi_\Sigma(\Phi^{-1}(u), \Phi^{-1}(v))$, and Φ is the CDF of a multivariate Gaussian distribution with correlation matrix $\left[\begin{smallmatrix} 1 & \theta \\ \theta & 1 \end{smallmatrix}\right]$.*

To see how the correlation parameter θ affects the dependency between the screening and hiring manager scores, consider the screening scores Q_s^S and hiring managers scores Q_s^H of the *shortlisted* candidates in Figure 1. The top row shows the joint distribution of Q_s^S and Q_s^H for different values of θ, and the bottom row shows the marginal distributions of Q_s^S and Q_s^H. First, note that the screening scores of the shortlisted candidates (Q_s^S shown in blue [//]) will no longer be uniformly distributed since the screening process selects the top scoring candidates according to the screener. In the left-most column, when $\theta = 0$, there is no correlation between the screening scores and the hiring manager's scores, so the marginal distribution of the hiring manager scores (Q_s^H shown in orange [\\]) reflects the underlying quality distribution, which is uniform. In the right-most column, when $\theta = 1$, the hiring manager scores are perfectly correlated with the screening scores, so the marginal distribution of the hiring manager scores follows the same distribution as the screening scores. In the middle, when $\theta = 0.5$, the hiring manager scores are partially correlated with the screening scores, so the marginal distribution of the hiring manager scores is somewhere between the underlying uniform distribution and the distribution of screening scores.

Figure 1: Dependency between screening and hiring manager scores – Gaussian Copula



3.1.2 Mixture Distribution Dependency

Although flexible, a drawback of the Gaussian copula is that it does not yield a tractable closed-form expression for the hiring manager score distribution Q_s^H preventing us from obtaining analytical solutions. However, we can reverse engineer a distribution for Q_s^H, such that it is tractable, reasonably approximates the Gaussian copula dependency, and has the desired properties we have seen previously. These properties are:

1. The distribution is parameterized by θ.
2. When $\theta = 1$, the distribution of Q_s^H is the same as the distribution of Q_s^S – i.e., $f_{Q_s^H} = f_{Q_s^S}$, when $\theta = 1$.
3. When $\theta = 0$, the distribution of Q_s^H is uniform. i.e. $f_{Q_s^H} = 1$, when $\theta = 0$.
4. When $0 < \theta < 1$, the distribution of Q_s^H is a "mix" of the uniform distribution and the distribution of Q_s^S.

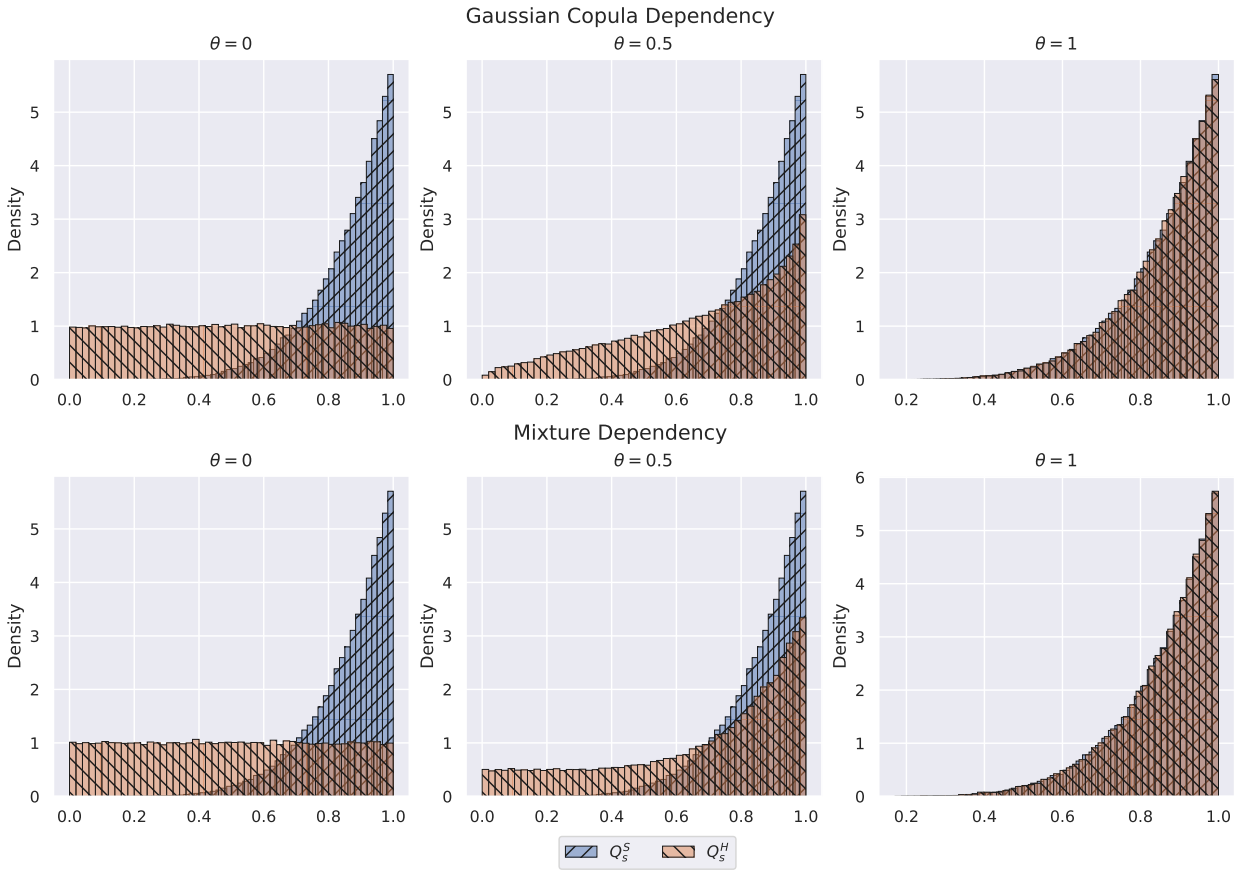
A mixture distribution between the distribution of Q_s^S and the uniform distribution with mixing parameter θ satisfies all the above properties. Under this dependency, given the density of the shortlist screening score $f_{Q_s^S}$ and the density of the uniform (which is 1), the density of the shortlist hiring manager score is:

$$f_{Q_s^H}(y;\theta) = \theta \cdot f_{Q_s^S}(y) + (1-\theta) f_U \tag{3.3}$$

$$= \theta \cdot f_{Q_s^S}(y) + (1-\theta) \tag{3.4}$$

As with the gaussian copula, when $\theta = 0$, the hiring manager score is uniform. When $\theta = 1$, the hiring manager score distribution is the same as the screening score. When $0 < \theta < 1$, the hiring manager score is a mixture of the two (See Figure 2).

Figure 2: Dependency between screening and hiring manager scores



4 Theoretical Results

We will now solve the 2-stage hiring model and provide an analytical solution for the probability that a female candidate is hired for a simple case where the size of the shortlist is 2 ($n_s = 2$), the number of hires is 1 ($n_h = 1$), and the dependency structure is the mixture distribution. For more general cases ($n_s > 2$, $n_h > 1$, other dependency structures), we show via simulations in Section 4.5 that the following results hold. Our three main theoretical results are that the probability of a female hire: (1) decreases with the correlation parameter, (2) increases with the size of the applicant pool, and (3) decreases with gender difference in correlation.

4.1 Perfect Observation of Candidate Quality – A Baseline

Assume, for a moment, that both the screening algorithm and the hiring manager can observe the true quality of the candidate perfectly – i.e., $Q = Q^S = Q^H$. Indeed, in such a case, we would not need a hiring manager in the first place but assume it's the case for expository purposes. Our goal here is to establish a baseline probability that a female is hired and show the line of reasoning.

Lemma 1. *The shortlisted candidate's true quality follows a beta distribution.*

Proof. Given the size of the shortlist $n_s = 2$, under parity constraint, the screening algorithm will shortlist the best male candidate and the best female candidate. Using the basics of order statistics, we know that the max of a sample of size n sampled from a uniform distribution between 0 and 1 is beta-distributed following the $Beta[n, 1]$ distribution. Given that there are $n_a \cdot p$ females in the applicant pool, whose quality scores are uniformly distributed, the shortlisted female candidate's screening score will have a beta distribution:

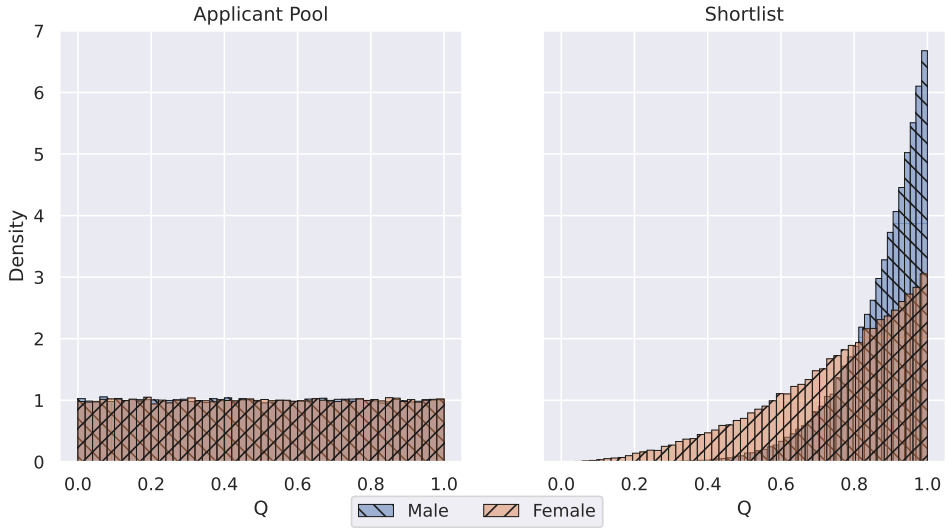
$$\begin{aligned} Q_{s,f} &\sim Beta[n_a p, 1] \\ f_{Q_{s,f}}(x) &= \frac{x^{-1+n_a p}}{Beta[n_a p, 1]} \end{aligned} \tag{4.1}$$

Similarly, the shortlisted male candidate's screening score distribution will be:

$$\begin{aligned} Q_{s,m} &\sim Beta[n_a(1-p), 1] \\ f_{Q_{s,m}}(x) &= \frac{x^{-1+n_a(1-p)}}{Beta[n_a(1-p), 1]} \end{aligned} \tag{4.2}$$

□

Figure 3: Q follows a uniform distribution in the applicant pool and beta distribution in the shortlist



Lemma 2. *The expected quality of the shortlisted candidate increases with the size of the applicant pool.*

Proof. The expected max of a sample increases with the size of the sample. Because the screening process shortlists the best candidates from an applicant pool whose quality scores are uniformly distributed between 0 and 1, the screening score of the shortlisted candidate will eventually reach 1 with increasing applicant pool size. The expected qualities of the shortlisted female and male candidates are:

$$\begin{aligned} E[Q_{s,f}] &= \frac{n_a p}{n_a p + 1} \\ E[Q_{s,m}] &= \frac{n_a (1-p)}{n_a (1-p) + 1} \end{aligned} \tag{4.3}$$

And the expected quality of the final hire is:

$$E[Q_h] = \frac{n_a}{n_a + 1} \tag{4.4}$$

□

Lemma 3. *Under parity constraint, the expected quality of the shortlisted male is greater than the expected quality of the shortlisted female.*

Proof. Lemma 2 shows that the expected quality increases with the applicant pool size. By definition, since males are the majority group, there will be more male applicants than female applicants in the applicant pool ($n_a(1-p) > n_a p$). This implies that the expected quality of the shortlisted male will be greater than the expected quality of the shortlisted female.

$$\begin{aligned} \frac{n_a p}{n_a p + 1} &< \frac{n_a(1-p)}{n_a(1-p)+1} \\ E[Q_{s,f}] &< E[Q_{s,m}] \end{aligned} \tag{4.5}$$

Note that this comparison is only in expectation. Since Q is a random variable, there will be cases where some shortlisted female candidates have higher quality scores than male candidates (see Figure 3). How often this happens is the subject of inquiry in what follows. □

Proposition 1. *When both the screening algorithm and the hiring manager can observe the candidate's quality perfectly, the gender parity constraint in the shortlist does not affect the gender proportion of hires.*

Proof. With no parity constraint, the probability that a female is shortlisted will be equal to the proportion of females in the applicant pool, as all candidates are drawn from an identical quality distribution.

With parity constraint, it is ensured that a female is shortlisted; however, the quality of the shortlisted candidates, conditional on being shortlisted, is no longer the same. The expected quality of the shortlisted male is higher than the shortlisted female (Lemma 3), meaning that the male candidate will be more likely to get hired from the shortlist.

Formally, consider the probability that a female will be hired. This happens when the probability

of the shortlisted female's quality exceeds the shortlisted male's quality.

$$
\begin{aligned}
Pr(\text{Female is hired}) &= Pr(Q_{s,f} > Q_{s,m}) \\
&= \int_0^1 Pr(Q_{s,m} = y) \cdot Pr(Q_{s,f} > y)\,\mathrm{d}y \\
&= \int_0^1 f_{Q_{s,m}}(y) \cdot (1 - F_{Q_{s,f}}(y))\,\mathrm{d}y \\
&= \int_0^1 \frac{y^{-1+n_a(1-p)}}{Beta[n_a(1-p),1]} \cdot (1 - \int_0^y \frac{z^{-1+n_a p}}{Beta[n_a p,1]}\,\mathrm{d}z)\,\mathrm{d}y \\
&= p
\end{aligned}
\tag{4.6}
$$

This probability is equal to the proportion of females in the applicant pool p. That is, when the candidate's quality can be observed perfectly, the probability that a female is hired is equal to the proportion of females in the applicant pool, regardless of whether there is a parity constraint.

□

4.2 Correlation between Screening and Hiring Manager Scores – θ

Now that we have established some baseline results, we can move on to a more realistic scenario, where the screening algorithm and the hiring manager observe a noisy estimate of the candidate's true quality. Because these estimates are noisy, the screening algorithm's scores and the hiring manager's scores no longer need to be the same. Instead, we model the dependency between the two scores using a mixture distribution parameterized by the parameter θ. As before, our goal is to find the probability that a female candidate is hired – so we will follow the same procedure as in the previous section. However, unlike before, the hiring manager's scores of the shortlisted candidate will no longer be beta distributed.

Lemma 4. *The shortlisted candidate's hiring manager score follows a mixture distribution that is a mixture of beta and uniform.*

Proof. Under the mixture distribution dependency structure, we show in Equation (3.3) that the hiring manager score follows a distribution that is a mixture of f_Q^S and uniform.

Since f_Q^S is beta distributed, the hiring manager's score of the shortlisted female candidate will be a mixture of beta and uniform:

$$f_{Q_{s,f}^H}(y) = \theta \frac{y^{-1+n_a p}}{Beta[n_a p, 1]} + (1-\theta) \tag{4.7}$$

And for the shortlisted male candidate:

$$f_{Q_{s,m}^H}(y) = \theta \frac{y^{-1+n_a (1-p)}}{Beta[n_a (1-p), 1]} + (1-\theta) \tag{4.8}$$

□

Lemma 5. *Under parity constraint, when $\theta > 0$, the expected quality of the shortlisted male is greater than the expected quality of the shortlisted female.*

Proof. Lemma 2 shows that the expected quality increases with the applicant pool size. By definition, since male is the majority group, there will be more male applicants than female applicants in the applicant pool ($n_a(1-p) > n_a p$). This implies that the expected quality of the shortlisted male will be greater than the expected quality of the shortlisted female.

$$\begin{gathered} \theta \frac{n_a p}{n_a p + 1} + (1-\theta)\frac{1}{2} < \theta \frac{n_a (1-p)}{n_a (1-p) + 1} + (1-\theta)\frac{1}{2}, \quad \theta > 0 \\ E[Q_{s,f}^H] < E[Q_{s,m}^H], \quad \theta > 0 \end{gathered} \tag{4.9}$$

□

Proposition 2. *The probability that a female is hired decreases when the correlation θ between screening and hiring manager scores increases.*

Proof. Consider the case when the screening and hiring manager scores are perfectly correlated ($\theta = 1$). This happens, for example, when the screener and the hiring manager use the exact same

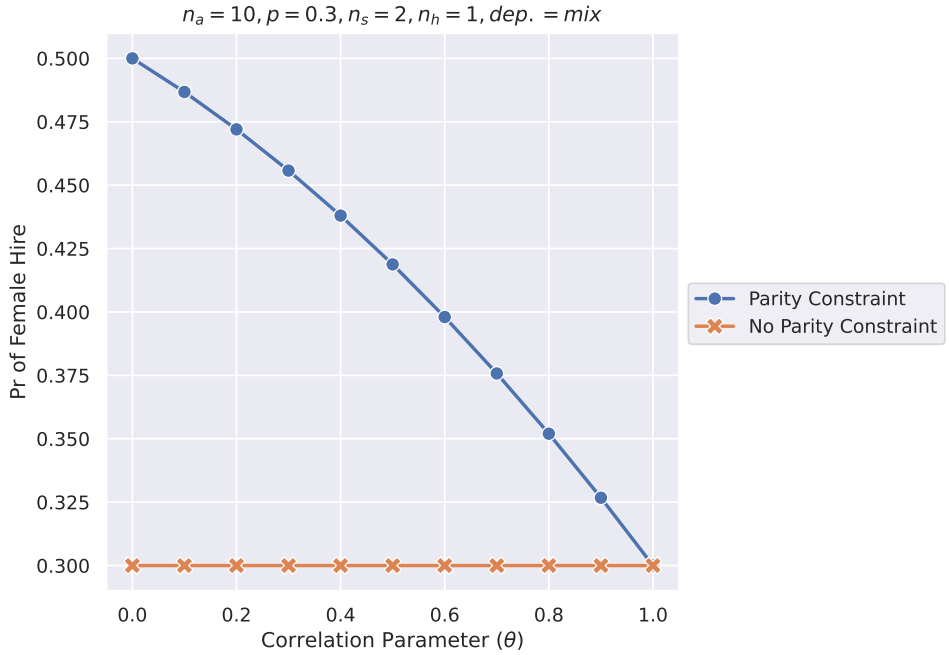
criteria to assess the quality of the candidate – for exposition, assume they both only look at the candidate's years of experience as an assessment of quality. After the screening stage, the shortlisted female candidate is less likely to have more years of experience than the male candidate (Lemma 5). If the hiring manager uses the exact same criteria, nothing changes, and the female candidate is still less likely to have more years of experience than the male candidate. Therefore, when $\theta = 1$, the shortlisted female candidate is less likely to be hired than the male candidate. This probability is equal to the proportion of women in the applicant pool p (See Appendix A). In other words, when screening and hiring manager scores are perfectly correlated, the parity constraint has no effect.

In the other extreme, consider a case when the two scores are perfectly uncorrelated ($\theta = 0$). Say, for exposition, that the screening algorithm only looks at the candidate's years of experience, and the hiring manager only looks at the candidate's college GPA (assuming they are indeed uncorrelated). After the screening stage, the shortlisted female candidate is less likely to have more years of experience than the male candidate. But since the hiring manager is only considering college GPA, both the shortlisted male and female candidates are equally likely to have a higher GPA (since the underlying quality distributions are the same). Therefore, when $\theta = 0$, the probability that a female is hired equals $\frac{1}{2}$. We present a proof in Appendix A, and show via simulations that the results hold for general cases ($n_s > 2$, $n_h > 1$, other dependency structures) in Section 4.5.

□

Corollary. *When screening and hiring manager scores are perfectly uncorrelated, the parity constraint effectively balances the gender proportion of hires. In contrast, parity constraint has no effect when the scores are perfectly correlated.*

Figure 4: Probability of Female Hire vs. Correlation Parameter θ (30% Female in the App. Pool)



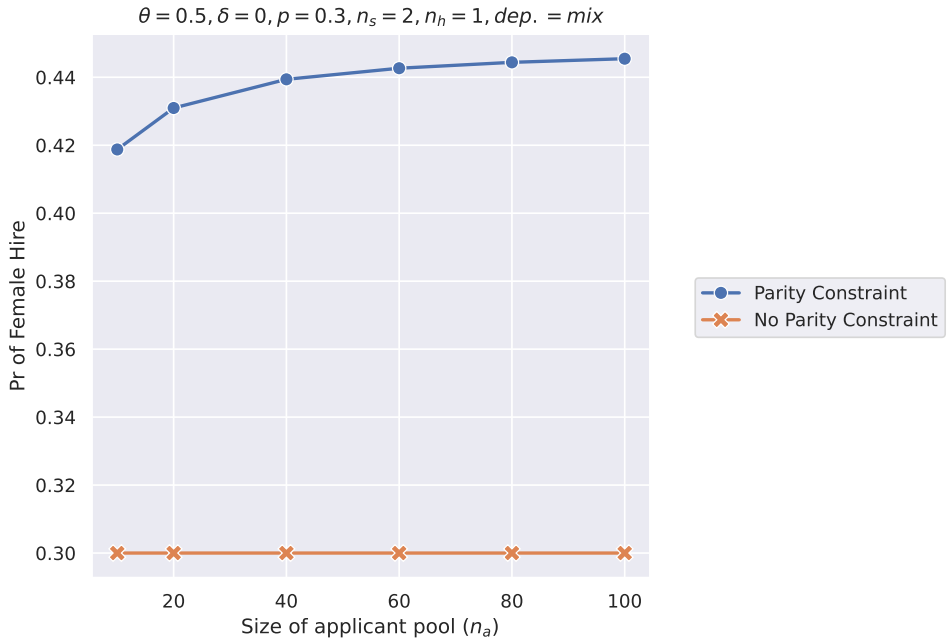
4.3 Size of the Applicant Pool

Proposition 3. *When $\theta < 1$, the probability that a female is hired increases with the size of the applicant pool.*

Proof. Lemma 2 shows that the expected quality increases with the size of the applicant pool. Since the quality score is bounded, the difference between the best male and best female candidate will get smaller as the overall size of the applicant pool increases. When there is a noise component to the screening score (i.e., when $\theta < 1$), any small differences will be overcome by the noise. We present a proof in Appendix A, and show via simulations that the results hold for general cases ($n_s > 2$, $n_h > 1$, other dependency structures) in Section 4.5.

□

Figure 5: Probability of Female Hire vs. Size of Applicant Pool (30% Female in the App. Pool)



4.4 The Case of a Biased Algorithmic Screener – δ

So far, we have assumed that the assessment criteria for men and women are the same. We now extend this model to allow for different assessment criteria using different correlation parameters, θ_m, and θ_f, for male and female candidates, respectively. Because we are fixing the assessment criteria of the hiring manager to be unbiased, any gender difference in the correlation parameter corresponds to differences in the screening algorithm's assessment – i.e., algorithmic bias. We parameterize this difference using δ, where:

$$\begin{aligned} \delta &= \theta_m - \theta_f \\ \theta_m &= \theta \end{aligned} \tag{4.10}$$

In a relevant practical example, the correlation parameters for men and women can be different when the screening algorithm is differentially predictive for men and women. Consider a screening algorithm that has no predictive power for female candidates' quality but high predictive power for male candidates. This algorithm will shortlist *random* female candidates such that the quality distribution of the shortlisted female candidates remains uniform. This implies that female candidates will have a low correlation with the hiring manager's score (See Figure 1). For male candidates, with

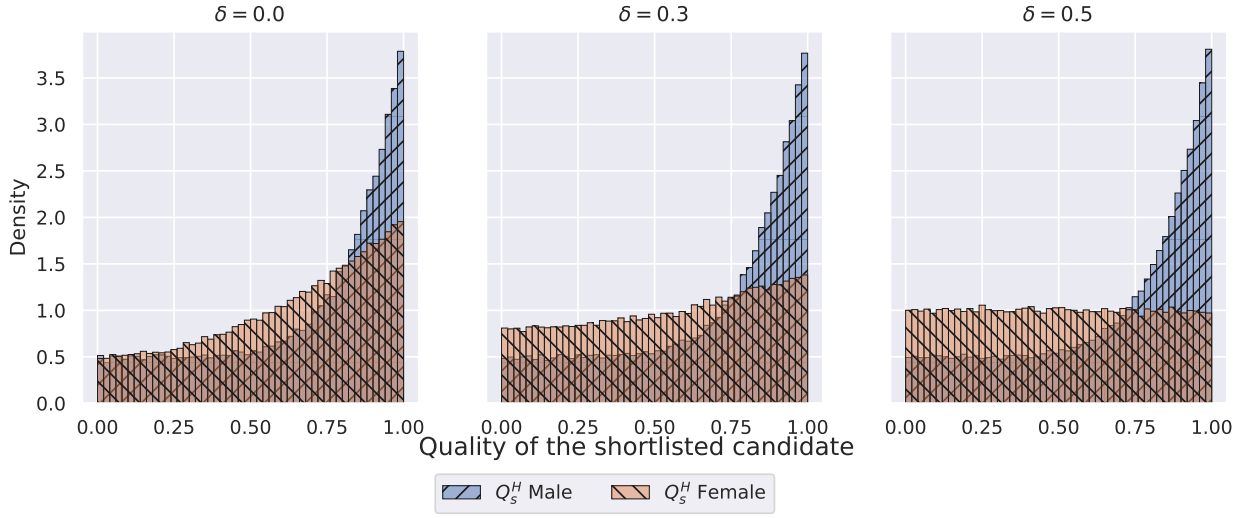
high predictive power, the algorithm will shortlist *"good"* male candidates such that the quality distribution of the shortlisted male candidates skews to the right. As a result, male candidates will have higher correlation with the hiring manager's score compared to female candidates – i.e., $\theta_m > \theta_f \implies \delta > 0$.

Therefore, when δ is positive, the screening algorithm is statistically biased against women. When δ is negative, the screening algorithm is statistically biased against men. When δ is zero, the screening algorithm is unbiased.

Proposition 4. *The probability that a female is hired decreases when the screening algorithm is less predictive for female candidates compared to male candidates.*

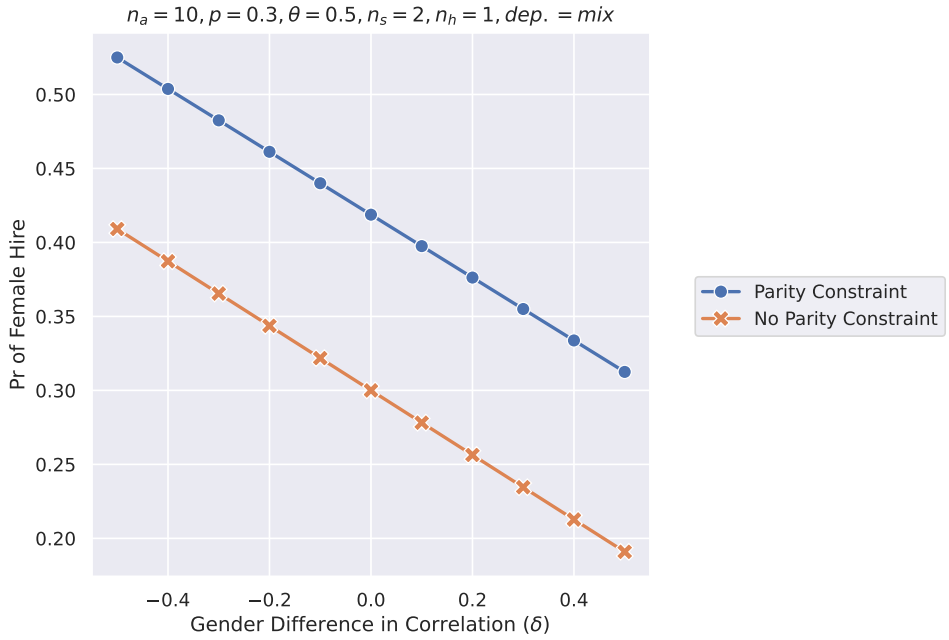
Proof. When the screening algorithm is less predictive for female candidates compared to male candidates, it implies that the correlation parameter for the female candidates will be less than that of male candidates, $\theta_f < \theta_m \implies \delta > 0$. In such a case, the hiring manager score distribution for the shortlisted female will be more uniform than the shortlisted male candidate, lowering her chances of being hired (See Figure 6, where, in the right-most panel, the screening algorithm is completely unpredictive for female candidates).

Figure 6: Quality of Shortlisted Candidate vs. δ



We solve for the probability that a female is hired as a function of δ and provide a solution in Appendix A, and simulation results in Section 4.5. □

Figure 7: Probability of Female Hire vs. Gender Difference in Predictive Power δ



4.5 Parametric Simulation Results

The previous section provides analytical solutions for the probability that a female candidate is hired as a function of model parameters for the case where $n_s = 2, n_h = 1$, *dependency* = *mixture*. To show that the main results also hold in the rest of the parameter space ($n_s > 2, n_h > 1$, *dependency* = *gaussian*), we run hiring simulations that simulate the data generating process of the underlying hiring model:

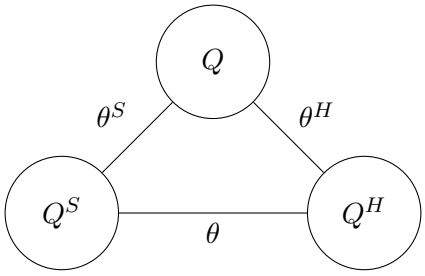
1. Applicant pool consists of $n_a p$ females and $n_a(1-p)$ males.
2. Both have the same uniform distribution of quality scores.
3. Under parity constraint, the screener shortlists the top $n_s/2$ male candidates and $n_s/2$ female candidates based on screening score Q^S.
4. Without parity constraint, the screener shortlists the top n_s candidates.
5. The hiring manager hires the best candidate from this shortlist based on the hiring manager score Q^H.

Given a list of model parameters $(\delta, \theta, n_a, p, n_s, n_h,$ *dependency*, *parity constraint*$)$ as inputs, the simulation algorithm will return the probability of a female hire (See Appendix B for the simulation algorithm).

We simulate a wide range of model parameters and report the results in Appendix C. In all cases, our main theoretical results – i.e., under parity constraint, the probability of a female hire decreases with θ and δ and increases with n_a – are replicated.

4.6 Expected Quality of Hires and the Cost of Parity Constraint

Parity constraint in the shortlist comes at the cost of candidate quality. Parity constraint necessarily implies that the expected screening score Q^S of candidates in the shortlist will be lesser than or equal to the case where there is no such constraint. How this loss of quality in the screening score Q^S translates to the loss in true quality Q in hires depends on the parameters discussed so far as well as two additional parameters: the correlation between the screening score and the true quality (θ^S) and the correlation between the hiring manager's score and the true quality (θ^H). θ^S measures the extent to which the screening score is predictive of the true quality (i.e., how good the screening algorithm is). θ^H measures the extent to which the hiring manager score is predictive of the true quality (i.e., how good the hiring manager is). And θ, as before, measures the correlation between the screening score and the hiring manager score.



We admit these additional parameters by extending the current model and jointly modeling Q, Q^S, Q^H with a 3-dimensional gaussian copula. Formally we model the joint distribution of Q, Q^S, Q^H as $F_{Q,Q^S,Q^H}(x, y, z; \Sigma) = C(F_Q(x), F_{Q^S}(y), F_{Q^H}(z))$, where C is the 3-dimensional gaussian copula.

$$C(u, v, k; \Sigma) = \Phi_\Sigma(\Phi^{-1}(u), \Phi^{-1}(v), \Phi^{-1}(k)) \tag{4.11}$$

$$\Sigma = \begin{bmatrix} 1 & \theta & \theta^S \\ \theta & 1 & \theta^H \\ \theta^S & \theta^H & 1 \end{bmatrix} \tag{4.12}$$

We define the cost of parity constraint as the difference in the expected true quality of the hire Q_{hire} with and without parity constraint.

$$\textit{Cost of Parity Constraint} = E[Q_{hire}]_{\textit{No Parity Constraint}} - E[Q_{hire}]_{\textit{Parity Constraint}} \tag{4.13}$$

We plot the expected quality of hire and the cost of parity constraint as a function of model parameters in Figure 8. For example, the top left panel shows the expected quality of hire as a function of θ^S conditioned on and averaged across all other parameters – i.e., $E_{\theta,\theta^H,\delta,n_a}[Q_{hire}]$ as a function of θ^S.

There are five main takeaways from these results.

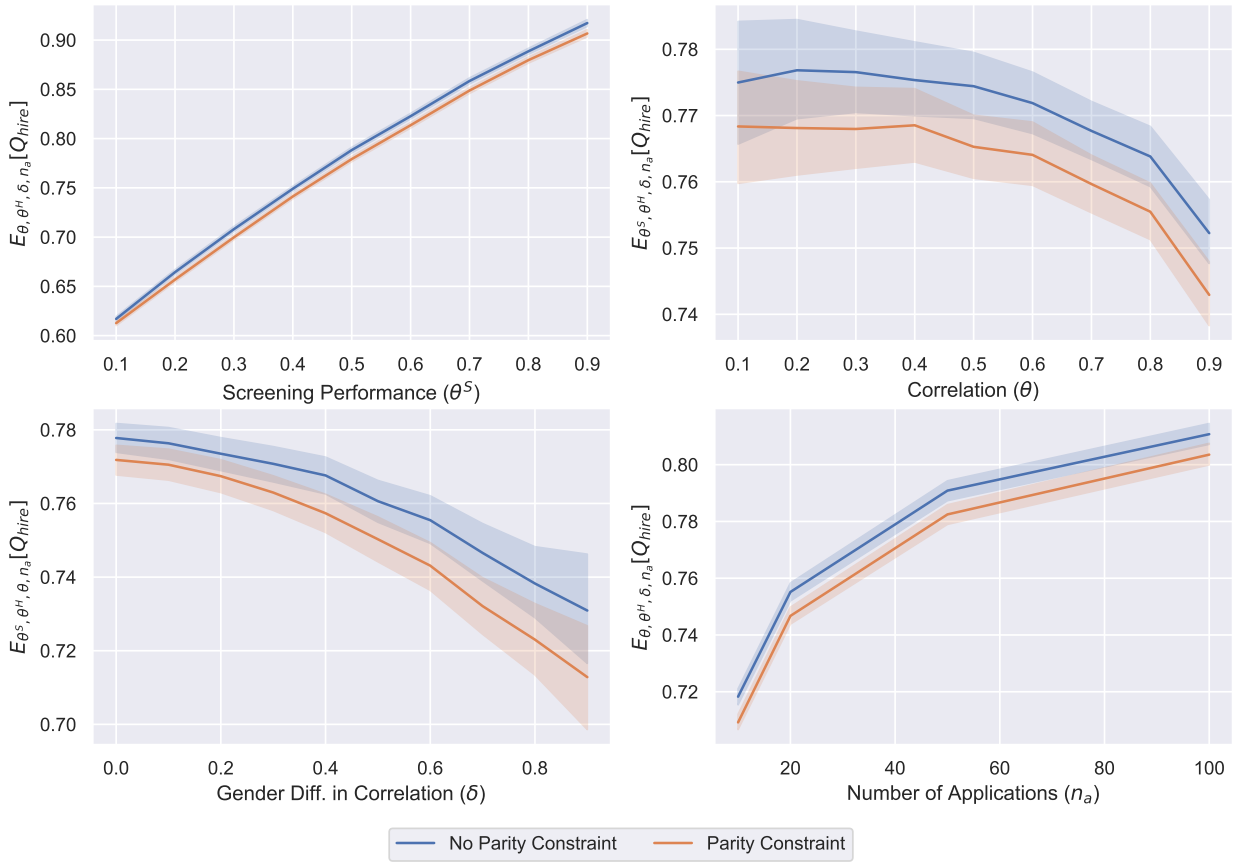
1. The cost of parity constraint is positive – i.e., the expected quality of hire under no parity constraint is always greater than orequal to the case when there is no parity constraint.

2. (Top Left) The expected quality of hire is increasing with θ^S – i.e., the better the screening score predicts true quality, larger the expected quality of hire.

3. (Top Right) The expected quality of hire decreases with θ. The 2-stage hiring process can be thought of as an aggregation of noisy signals from the screener and the hiring manager. When aggregating noisy signals, uncorrelated signals reveal more information about the true quality of the candidate than correlated signals (see e.g., Clemen and Winkler 1985). Hence, conditional on the informativeness of individual signals – i.e., conditional on θ^S and θ^H – the expected quality of hire decreases with θ.

4. (Bottom Left) The expected quality of hire decreases with δ. When $\delta > 0$, it implies that the screener is less predictive of Q^H for female candidates than male candidates ($\theta_f < \theta_m$).

Figure 8: Expected Quality of Hire



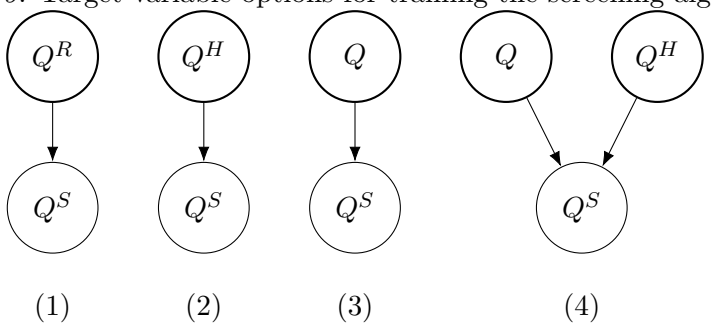
For the reasons stated in the previous point, the lower θ would increase the expected quality of hire, but this is only true *if* both male and female shortlists were proportionally equally likely to be hired by the hiring manager. However, since the hiring manager's likelihood of choosing the female candidate decreases with δ (Proposition 4), the expected quality of the hire also decreases. In other words, the would-be gain from the uncorrelated signal from female candidates does not contribute to the expected quality of hire, since they are less likely to be hired by the hiring manager.

5. (Bottom Right) The expected quality of hire increases with the size of the applicant pool. This is because the 2-stage hiring process selects the best from a pool of applicants, and the max of a sample increases with sample size.

4.7 Discussion and Implications for Algorithmic Design

So far we have taken the screening algorithm as a black-box model that computes a screening score Q^S given some job and candidate characteristics. However, we have control over how Q^S is computed through the design of the screening algorithm. The theoretical analyses provide insights for designing the optimal screening algorithms that maximizes both the expected quality of hire and the effectiveness of parity constraint (i.e., diversity). As we have shown: (1) the expected quality of hire is increasing with θ^S, (2) the probability of a female hire decreases with θ under parity constraint, and (3) conditional on θ^S, the expected quality of hire decreases with θ. This implies that to maximize hire quality *and* the effectiveness of parity constraint, we should maximize θ^S and minimize θ. In other words, we want the screening algorithm to be good at predicting true quality, but different from the hiring manager's assessment – i.e., we want the screening algorithm to be *complementary* to the hiring manager. To show how we can achieve this, we first consider all the available design options for training the screening algorithm.

Figure 9: Target variable options for training the screening algorithm



(1) The first option is to train the screening algorithm with the historical recruiter's scores/decisions Q^R as the target variable (i.e., decisions of human screeners that used to do the job of the algorithmic screener). The advantage with this approach is that the number of observations to train the algorithm is large and uncensored (i.e., the firm observes all applicants that applied and the recruiter's decision for each applicant). The disadvantages are that (a) recruiter's decision is only a proxy of true quality, so θ^S may not be maximal, and (b) we have no control over the effectiveness of the parity constraint, as it partly depends on how similar the recruiter's and hiring manager's assessment criteria are (i.e., we have no control over θ).

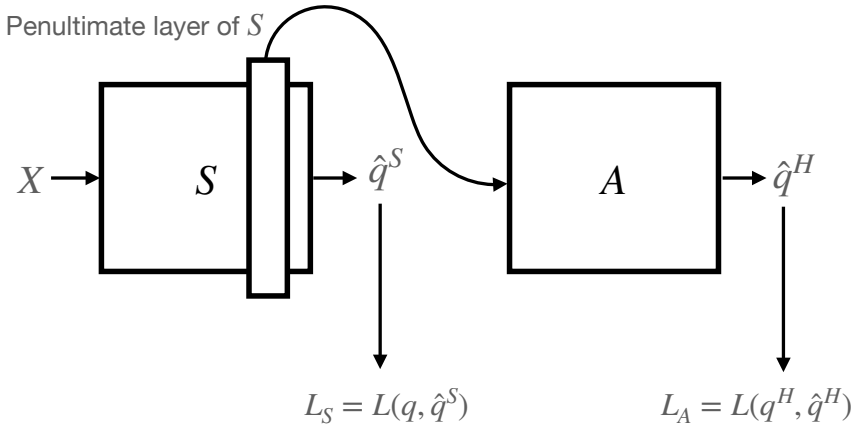
(2) The second option is to train the algorithm with hiring manager's scores/decisions Q^H as

the target variable. The parity constraint will not be very effective in this case, since the algorithm is designed to be similar to the hiring manager – i.e., θ is maximized by design.

(3) The third option is to the train the algorithm with true quality Q (e.g., job performance) as the target variable. Although this will maximize the screening performance in predicting quality (i.e., maximizes θ^S), we will have no control over the effectiveness of the parity constraint, since we have no control over θ.

(4) The last option is to combine the above two approaches and train the algorithm with both true quality Q and hiring manager's scores/decisions Q^H as the target variables. As we have argued, to optimize both the expected quality of hire and the effectiveness of parity constraint, we need the screening algorithm to be complementary to the hiring manager – i.e., we need to maximize θ^S and minimize θ. One way to do this is with adversarial learning (Zhang et al. 2018). In this model, the screening algorithm would consist of two components: (a) a predictor S that is trained to predict the candidate's true quality, and (b) an adversary A that tries to predict the hiring manager's estimate of quality. The overall model trades off the two objectives to learn a model that is good at predicting the candidate's true quality but poor at predicting the hiring manager's estimate of quality. More specifically, the predictor S takes as input candidate and job characteristics X and is trained to predict the candidate's true quality Q by minimizing loss $L_S = L(Q, \hat{Q}^S)$, which relates to θ^S. The adversary A takes as input the penultimate layer of A and is trained to predict the hiring manager's estimate of quality Q^H by minimizing loss $L_A = L(Q^H, \hat{Q}^H)$ (which relates θ) Both the predictor and the adversary are trained jointly by minimizing the overall loss $L = L_S - \lambda L_A$, where λ is a hyperparameter that controls the trade-off between minimizing θ and maximizing θ^S. The overall goal is to learn a representation for Q^S that predicts Q well (maximize θ^S) but Q^H poorly (minimize θ).

Figure 10: Illustration of Adversarial Learning for Screening Algorithm



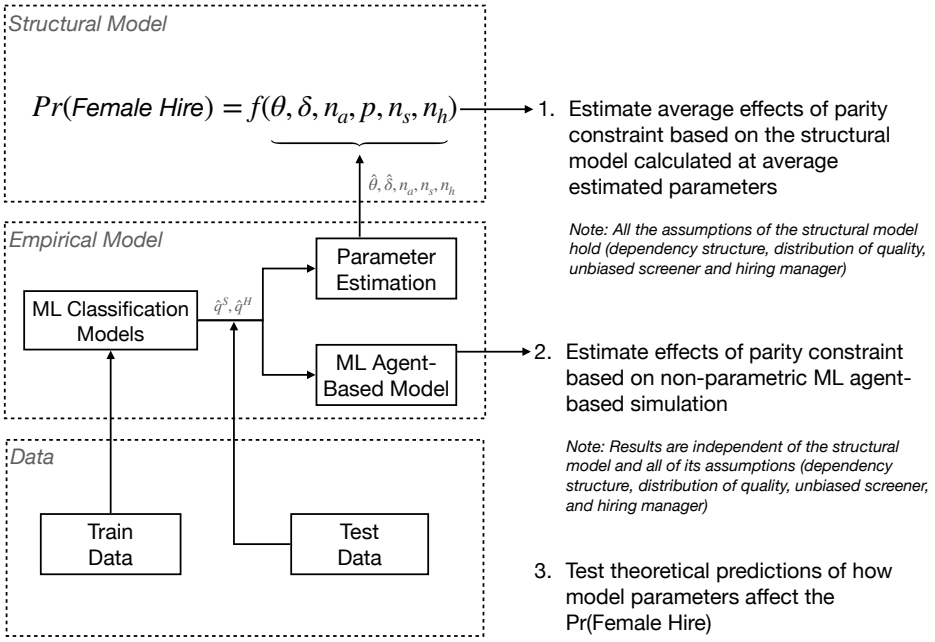
5 Empirical Modeling

Theoretical analyses in the previous sections show how the effectiveness of the parity constraint depends on key parameters, which in turn inform us how to design the optimal screening algorithm. Although we have shown that it is optimal to train the screening algorithm on both true quality and hire manager's estimate of quality, it's not typically what's done in practice. Typically, screening algorithms are trained on historical decisions of human screeners/recruiters because the training data is abundant and readily available. For example, *Linkedin Recruiter's* recommendation algorithm is trained on recruiter's decisions, because LinkedIn does not observe outcomes further down the hiring pipeline (Geyik et al. 2019). Similarly, firms may choose to train their screening algorithm on recruiter's decisions rather than true quality (e.g., job performance) if the training data for the latter is scarce[10]. If a firm were to train the screening algorithm on recruiter's decisions and implement a parity constraint, how effective the parity constraint would be in increasing diversity is an empirical question. The effectiveness of the parity constraint depends on the values of the parameters discussed thus far, which in turn depends on how screeners and hiring managers make hiring decisions in the real world, how correlated their assessment criteria are, how well the resume screening algorithm performs, and how large the applicant pool is – all of which are empirical questions.

[10]Firms observe recruiter decisions for all applicants, whereas they observe job performance only for candidates that were hired. Within our data, we observe that the hire-rate is between 1-5% across firms, meaning that the training data to predict recruiter decisions will be 20-100x larger than training data to predict job performance.

In this section, we introduce real hiring data into the model to estimate the effectiveness of the parity constraint using counterfactual policy evaluation. Specifically we ask, *if firms were to implement a parity constraint on a screening algorithm trained on recruiter decisions, how effective would it be in increasing diversity?* We address this question using two different approaches as illustrated in Figure 11: (1) parametric estimation and (2) ML agent-based simulation.

Figure 11: Illustration of Empirical Strategy



Our first approach is to estimate the point values of model parameters using hiring decisions and plug the estimated values into the structural model to evaluate the effectiveness of the parity constraint. Some of the model parameters, such as the size of the applicant pool n_a, size of the shortlist n_s, etc., are readily observed in the hiring data. In contrast, others (θ, δ) need to be estimated from the hiring decisions of screeners and hiring managers. To estimate θ and δ, we need screening and hiring manager scores (q^S, q^H) for each applicant, which are latent variables we can recover from binary decisions made by screeners and hiring managers. To recover these scores, we use two machine learning (ML) classification models trained on prior screening and hiring decisions to predict the screening and hiring manager scores, respectively. Overall, this approach allows us to estimate the structural model with and without parity constraint in the subset of the parameter space that is practically relevant to real-world hiring.

The second approach is an ML agent-based simulation approach that does not rely on the structural model and is fully non-parametric. Instead of estimating the model parameters and plugging them back into the structural model, we only rely on the ML classification models trained on prior screening and hiring decisions. Using these ML models, we create ML agents that are "digital twins" of screeners and hiring managers, which are then used to simulate screening and hiring decisions with and without the parity constraint. This approach does not rely on the structural model and is therefore independent of all of its assumptions (dependency structure, distribution of quality, unbiasedness of the hiring manager). The parametric estimation method is more faithful to the structural model than the data. In contrast, the ML agent-based estimation method is more faithful to the data than the structural model.

Once we estimate the model parameters and the effects of the parity constraint, we test the theoretical predictions of how the model parameters affect the probability of female hire by exploiting the variation in parameters across job postings.

To be clear, none of the companies in our dataset had a parity constraint policy. However, we can still estimate the effectiveness of the parity constraint and test the theoretical predictions even in the absence of such a policy because the model parameters are artifacts of how screeners and hiring managers make hiring decisions[11]. To summarize, our goals in this section are to:

1. Estimate the effects of parity constraint using parametric estimation of model parameters
2. Estimate the effects of parity constraint using ML agent-based simulation
3. Test the theoretical predictions of how model parameters affect the probability of a female hire by exploiting variation in parameters across jobs

5.1 Data

We use Applicant Tracking System (ATS) data from eight tech firms based in the U.S. These firms are clients of an HR analytics software provider, who provided us with the aggregated ATS data. The ATS keeps a detailed record of all the applicants, their characteristics (gender, years of experience, etc.), the applicant's resume, the job posting to which they applied and the corresponding business

[11]The assumption here is that the model parameters are plausibly exogenous to the parity constraint – i.e., screener's and hiring manager's assessment criteria do not change after instituting the parity constraint. This assumption is automatically satisfied for the algorithmic screener since it'll be trained on past decisions of a human screener. We discuss the plausibility of the hiring manager's assessment criteria changing in Section 7.

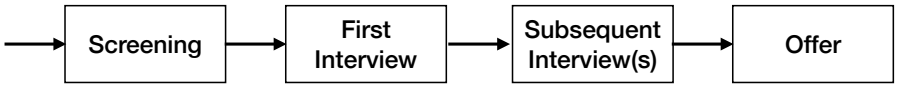
unit, and the application outcome (0/1) in each stage of the hiring funnel. We have 799k applicants (60% male, 40% female) across 3,608 job postings. We only consider external applicants (i.e., outside applicants who applied for a job) and disregard internal applicants (i.e., applicants who are already employed at the firm) and prospects/leads (i.e., candidates who are invited by recruiters to apply). We manually categorize each job posting into the following categories based on the business unit of the job.

Table 2: Number of Applicants and Job Postings by Job Category

Job Category	N Applicants	N Jobs
Engineering & Technical	214,943	1,178
Product & Design	130,669	534
Sales & Marketing	92,559	391
Legal & PR	75,955	332
Other	70,864	53
Finance & Accounting	69,536	308
Biz Dev & Operations	51,523	299
HR	48,122	246
Customer Service & Acct Management	42,199	238
Overall	799,108	3,608

Each firm may have slightly different hiring processes, but because we only consider external applicants, they all fit the following canonical hiring process: Screening, First Interview, Subsequent Interviews, and Offer. The first stage is the screening stage, where human screeners review the applications received. If candidates pass the screening stage, they move on to the first interview. If they clear that stage, they move on to subsequent interviews and finally to the offer stage.

Figure 12: Hiring Funnel



For an average job, 233 applicants apply, of which 36 pass the initial screening and are shortlisted for the first interview. After the first interview, 7 finalists move on to subsequent interviews, of which 2 receive an offer (there can be multiple vacancies per job posting).

Although this multi-staged hiring process deviates from our 2-stage hiring model, it is sufficient

to only examine the following two stages: (1) the stage with the parity constraint (Screening) and (2) the stage after the parity constraint (Interview 1). To see why this is the case, consider an example where the applicant pool is 70% male and 30% female. With a parity constraint, the proportion of male/female candidates after the screening stage will be 50/50. In Interview 1, the hiring manager can "undo" some of this parity constraint resulting in 60/40, for example. As long as the selection process in the subsequent stages is unbiased (which is the case in our model), the same proportion of 60/40 will be maintained in the rest of the hiring funnel.

5.2 Parametric Estimation

Our first goal is to estimate model parameters for each job using hiring decisions. The applicant pool size n_a, shortlist n_s, and finalist n_h do not require estimation – they are readily observed in the data. The correlation parameter θ and the gender difference in correlation δ require estimation.

The correlation parameter in the theoretical model is meant to capture the correlation in the assessment criteria of the screening algorithm (which will be trained on the decisions of human screeners) and the hiring manager. One way to crudely estimate this correlation might be to regress the observed screening decisions against hiring manager decisions for each applicant. The problem with this approach is two-fold: (1) we only observe binary decisions (0/1) of the screener and the hiring manager, and (2) we only observe hiring manager decisions for applicants who passed screening. This means we won't have any variation in screening decisions (all 1s), making this approach impossible. Instead, we need scores from the screener and hiring manager for *all* applicants (even those that did not pass screening) to calculate the correlation between the two.

5.2.1 Estimation of Screening and Hiring Manager Scores

We can exploit the variation in the screener and hiring manager's binary decisions and recover the latent screening and hiring manager scores, respectively. Formally, consider the binary choice

decision $y_{i,j}$ faced by the screener/hiring manager when assessing candidate i for job j.

$$y_{i,j} = \begin{cases} 1 & q > 0 \\ 0 & q \leq 0 \end{cases} \tag{5.1}$$

$$q_{i,j} = f(X_i, X_j, \epsilon_{i,j}) \tag{5.2}$$

The decision maker makes a positive decision $y_{i,j} = 1$ if the net utility from applicant i exceeds 0 ($q > 0$), and makes a negative decision otherwise. The net utility is a function f of applicant characteristics X_i, job characteristics X_j as well as some idiosyncratic error $\epsilon_{i,j}$. In the hiring data, we observe the binary decisions $y_{i,j}$, applicant characteristics X_i (resume text), and job characteristics X_j (job description text) for each applicant-job pair. We can use this information and train two ML models, one for the screener and one for the hiring manager, to flexibly estimate f^S and f^H, using which we can estimate screening and hiring manager scores. The screening model takes in an applicant's resume and job description texts as inputs and predicts the probability of being shortlisted ($\hat{f}^S : (X_i, X_j) \mapsto \hat{y}^S$). The hiring manager model takes in an applicant's resume and job description texts as inputs and predicts the applicant's probability of passing the interview ($\hat{f}^H : (X_i, X_j) \mapsto \hat{y}^H$). We discuss these models in detail in Section 5.3.

To estimate the screening score $\hat{q}^S_{i,j}$, we take the quantile score[12] of the predicted probability of being shortlisted for applicant i in job posting j. Formally, let $\hat{y}^S_{i,j}$ be the predicted probability of being shortlisted for candidate i in job posting j, and let $\hat{\boldsymbol{y}}^S_j$ be the list of predicted probabilities for all candidates in job posting j. The screening score $\hat{q}^S_{i,j}$, and similarly, hiring manager score $\hat{q}^H_{i,j}$ for candidate i in job posting j are given by:

$$\hat{q}^S_{i,j} = Quantile(\hat{y}^S_{i,j}, \hat{\boldsymbol{y}}^S_j) \tag{5.3}$$

$$\hat{q}^H_{i,j} = Quantile(\hat{y}^H_{i,j}, \hat{\boldsymbol{y}}^H_j) \tag{5.4}$$

[12]The quantile transformation is analogous to the inverse integral transform in our theoretical model. It maps the scores to a uniform distribution

5.2.2 Estimation of θ

Once we have the screening and hiring manager scores for each candidate, we can estimate the correlation parameter $\hat{\theta}_j$ for job posting j in the test set using the Pearson correlation coefficient. The Pearson correlation coefficient equals the correlation parameter θ in our structural model with Gaussian copula dependency. We calculate the average by taking a weighted sum of the parameters (weighted by the size of applicant pool n_a) across job postings.

$$\hat{\theta}_j = \rho_{\hat{q}^S, \hat{q}^H} = \frac{Cov(\hat{q}^S, \hat{q}^H)}{\sigma_{\hat{q}^S} \sigma_{\hat{q}^H}} \tag{5.5}$$

$$\hat{\theta}_{avg} = \sum_j \frac{n_{a,j}}{n_a} \hat{\theta}_j \tag{5.6}$$

5.2.3 Estimation of δ

To get an empirical estimate of δ, we calculate $\hat{\theta}$ for male and female applicants separately for each job posting j in the hold-out test set and take the difference.

$$\hat{\delta}_j = \hat{\theta}_{j,m} - \hat{\theta}_{j,f} \tag{5.7}$$

$$\hat{\delta}_{avg} = \sum_j \frac{n_{a,j}}{n_a} \hat{\delta}_j \tag{5.8}$$

5.3 Building ML Models for Screening and Hiring

To train our resume screening and hiring manager models, we use a state-of-the-art deep learning model for natural language processing called BigBird[13] (Zaheer et al. 2020). BigBird uses a variant of the popular BERT-style transformer architecture and is optimized for long documents[14] (Vaswani et al. 2017; Devlin et al. 2019).

For both the screening and hiring manager model, we use an 80/10/10 train/evaluation/test split, where we use the evaluation set for hyperparameter tuning and the hold test set for final model evaluation, prediction, and parameter estimation. We follow Sun et al. (2019) for initial hyperparameters and fine-tune them by making small adjustments. The following parameters yield

[13] https://github.com/google-research/bigbird

[14] The classic BERT architecture uses a self-attention mechanism that scales quadratically with the number of tokens in the document. This makes using BERT for long documents such as resumes infeasible due to memory and computational footprint. BigBird overcomes this by using a sparse attention mechanism that scales linearly with the number of tokens in the document.

the best results based on the area under ROC criteria on the evaluation set: `Epochs=3, Batch Size=14, Learning Rate=2e-5, Weight Decay=2e-5`.

5.3.1 Data Preparation

The outcome of an application depends on both the applicant's characteristics and the job characteristics – i.e., it depends on the match between the applicant and the job to which they applied. Therefore, the input data should contain both the characteristics of the applicant (resume) and the job (job description).

One way to include both the applicant and the job characteristics is to concatenate the resume and job description text together and feed the concatenated text as a single document to the model. A drawback, however, is that job descriptions tend to be long and full of boilerplate language that does not contain any signal about the outcome of an application. So using the full job description increases the length of each document, which puts unnecessary memory and computational strain on training the model. To overcome this, we get the most important characteristics from the job (company name, job name, business unit, employment type, location, skills, and keywords), concatenate these characteristics with the resume text and source of the application, and feed the concatenated text as a single document to the model. We get the company name, job name, business unit, and location directly from the ATS. For skills and keywords, we use a dictionary of skills that was created in a separate analysis by aggregating all the skills and keywords listed in the Skills section of LinkedIn profiles. We then concatenate this text with the resume text to create a single document (See Figure 13). The model parses this document into tokens, embeds each token into a vector representation, and creates a tensor representation (an n-dimensional matrix) for the document, which is then fed into a deep neural network.

Figure 13: Sample Input Instance

```
company x
job_name=
backend engineer
biz_unit=
infrastructure
job_loc=
san francisco, ca
job_skills=
build tools, full stack, web development,
impact investing, c, shell, relational
databases, big data, debugging, design,
mobile, python, unix, sql, software
engineering, ruby…
employment_type=
fulltime
source=
jobsite
resume=
john doe
123 center st. new york, ny
education
b.s computer science nyu, ny - may 2015
gpa: 3.6/4
relevant coursework: database design,
operating systems
…
experience
software engineering intern, company y, summer
2013
```

5.3.2 Train/Test Split – Stratification on Job Postings

A simple approach for training and evaluating a model is to randomly split the data into train and test sets and evaluate on the test set. However, such an approach would inflate the model performance metrics since the model would have already seen the applicants and outcomes within a given job posting. A company trying to train a resume screening model would not take such an approach because what matters is how well the model performs on new, unseen job postings. We consider this and split the dataset by stratifying on job postings. We randomly take 80% of the job postings for the training set, 10% for the evaluation set, and 10% for the hold-out test set. This ensures that the model is evaluated on a test set containing applicants and job postings that the model has never seen before.

For the screening model, the size of the training/evaluation set is 725,351, and the hold-out test set is 73,757. For training the hiring manager model, we only use the subset of candidates

that actually got shortlisted (because we don't know the counterfactual interview outcome for the candidates that did not get shortlisted). Therefore, for the hiring manager model, the size of the training/evaluation set is 106,419, and the hold-out test set (for performance evaluation) is 11,357. Note, however, that we use the full hold-out test set for estimation of hiring manager scores $\hat{q}^H$ and model parameters θ and δ.

5.4 ML Agent-Based Estimation

Our second approach for estimating the effects of parity constraint is based on ML agent-based simulation. Here, the screening and hiring manager ML models serve as the agents who make screening and interview decisions using the decision rules as described below (See Appendix B for a detailed procedure).

1. Within job posting j, get the size of the shortlist n_s and the size of the finalist n_h.
2. For each applicant i in j, estimate the screening score $\hat{q}^S_{i,j}$ using the ML screening model.
3. Rank order the applicants by screening score $\hat{q}^S_{i,j}$.
4. Under parity constraint, shortlist the top $n_s/2$ male candidates and the top $n_s/2$ female candidates.
5. Without parity constraint, shortlist the top n_s candidates.
6. For each shortlisted candidate, estimate the hiring manager score $\hat{q}^H_{i,j}$ using the hiring manager ML model.
7. Rank order the shortlisted candidates by hiring manager score $\hat{q}^H_{i,j}$, and select the top n_h candidates.
8. The estimated Pr(Female Hire) is the proportion of female candidates in the finalist.

An advantage of this approach is that does not rely on the structural model and is therefore independent of all of its assumptions (dependency structure, distribution of quality, unbiasedness of the hiring manager). This allows us to run more realistic counterfactual simulations that is grounded in data.

5.5 Test of Theoretical Predictions

Lastly, our third goal in this section is to test the theoretical predictions of how model parameters affect the probability of a female hire by incorporating hiring data. We show through our theoretical

model and simulation that under parity constraint, the probability of a female hire decreases with parameters θ and δ and increases with the size of the applicant pool n_a. Although we tested these results using a wide range of parameters, these results are still subject to the parametric assumptions about the quality distributions of the applicants, dependency structure between the screener and the hiring manager, etc. We aim to overcome this limitation by testing our theoretical predictions using hiring data, where these parametric assumptions do not necessarily hold.

We test the theoretical predictions using the simplest possible model (a logit model), where the dependent variable is the estimated probability of female hire, the independent variables are the main model parameters (θ, δ, n_a), and the controls are the proportion of women in the applicant pool p, size of the shortlist n_s, and size of finalists n_h – all observed/estimated at the job posting level j. We are able to estimate this model using the variation in these variables across job postings.

$$logit(\hat{Pr}(FemHire)_j) = \beta_0 + \beta_1\hat{\theta}_j + \beta_2\hat{\theta}_j^2 + \beta_3\hat{\delta}_j + \beta_4\hat{\delta}_j^2 + \beta_5 n_{aj} + \beta_6 n_{aj}^2 + \beta_7 p_j + \beta_8 n_{sj} + \beta_9 n_{hj} + \epsilon_j \tag{5.9}$$

For $\hat{\theta}_j$ and $\hat{\delta}_j$, we use the estimated model parameters for each job opening j as outlined in Section 5.2. For variables n_a, p, n_s, n_h, we use the observed values in the data. For the outcome, $\hat{Pr}(FemHire)$, we estimate the probability of a female candidate being hired using the ML agent-based simulation described in the previous section. We add squared terms of the main model parameters to capture non-linear effects as predicted by the theoretical model.

6 Empirical Results

This section reports the results of the empirical models. All of the following results are estimated on the hold-out test set.

6.1 Predictive Performance of ML Models

We measure the predictive performance of the ML models using the Area Under ROC curve (AUC) criteria[15] on the hold-out test set and report the results in Table 3.

[15] AUC is a widely-used measure for predictive performance for classification models since it's agnostic to both imbalanced classes and classification thresholds. The score ranges from 0.5 to 1, where 0.5 corresponds to a random classifier, and 1 corresponds to a perfect classifier.

For the screening model, the overall AUC score is 0.83, and there is no difference in AUC scores between the male and female candidates (see Appendix D for the ROC curves). We also find that there is some heterogeneity in performance across job types, as reported in Table 4.

For the hiring manager model, the predictive performance is lower compared to the screening model since the hiring manager has more information from the interview, which we do not observe. Nonetheless, the predictive performance based on just resume characteristics is still reasonably high at 0.72, and there is no difference between genders. Note that the hiring manager model is evaluated on a subset of applicants in the hold-out test set who were, in fact, shortlisted.

To qualitatively assess the models and to understand the assessment criteria used by the screener and the hiring manager, we recover the most predictive features from the screening and hiring manager models using SHAP values and report the results in Appendix D.2.

Table 3: Predictive Model Performance by Gender on Hold-out Test Set

	Screening		Interview	
Group	AUC	Support	AUC	Support
Female	0.83	31,364	0.72	4,679
Male	0.83	42,393	0.72	6,678
Overall	0.83	73,757	0.72	11,357

Table 4: Predictive Model Performance by Job Category on Hold-out Test Set

	Screening		Interview	
Job Category	AUC	Support	AUC	Support
Legal & PR	0.86	7,337	0.73	926
Product & Design	0.85	12,519	0.71	1,863
Sales & Marketing	0.85	15,169	0.68	2,120
Other	0.83	214	0.33	25
Engineering & Technical	0.82	16,506	0.71	3,315
Finance & Accounting	0.82	7,026	0.76	886
Biz Dev & Operations	0.81	4,836	0.7	628
HR	0.79	3,355	0.73	452
Customer Service & Acct Management	0.78	6,734	0.72	1,112
Overall	0.83	73,757	0.72	11,357

6.2 Parameter Estimates

We estimate the model parameters (θ, δ, n_a) for each job posting separately and plot the distribution in Figure 14. For the estimation of θ and δ, we follow the procedure described in Section 5.2. For n_a, we report the observed size of the applicant pool. We also aggregate the parameters at the job category level by taking the weighted average across job postings and report the results in Table 5. Note that these parameter estimates are calculated on the hold-out test set.

The average estimate of $\hat{\theta}$ is 0.56, with a high level of heterogeneity across jobs. The average estimate of $\hat{\delta}$ is close to 0, implying that the screening algorithm is statistically unbiased. This is in line with measures of the screening algorithm's predictive power (AUC), which is also equal between male and female candidates.

Figure 14: Distribution of Parameter Estimates across Job Postings

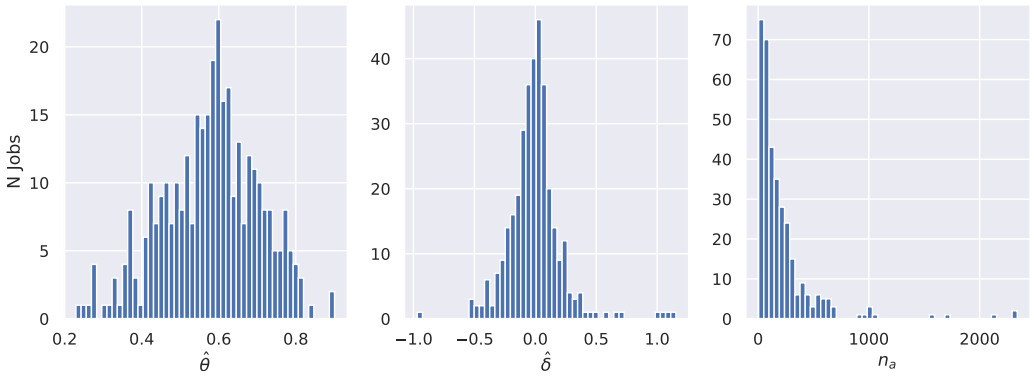


Table 5: Average Parameter Estimates

Job Category	$\hat{\theta}$	$\hat{\delta}$	n_a
Finance & Accounting	0.593	-0.009	184
Sales & Marketing	0.592	0.018	329
Customer Service & Acct Management	0.58	-0.011	240
Engineering & Technical	0.562	0.007	160
Biz Dev & Operations	0.544	0.015	156
Product & Design	0.528	0.031	215
HR	0.515	-0.032	159
Legal & PR	0.509	-0.016	229
Other	0.236	0.035	107
Average	0.558	0.007	205

6.3 Effectiveness of Parity Constraint

We estimate the overall effect of the parity constraint on the gender proportion of hires using counterfactual policy evaluation. We do this in two ways: (1) parametric estimation and (2) ML agent-based simulation.

6.3.1 Parametric Estimation

We estimate the effectiveness of the parity constraint by taking the average estimates of model parameters from the previous section and inputting them into the structural model using the procedure described in Section 4.5. We set the proportion of females in the applicant pool to be $p = 0.3$. Here, all the assumptions of the structural model hold (same quality distribution among men and women, unbiased hiring manager, dependency structure).

Table 6: Effectivess of Parity Constraint – Parametric Estimation

							Prop. of Females		
Dependency	Parity Constraint	n_a	n_s	n_h	θ	δ	App. Pool	Shortlist	Finalist
Mixture	False	233	36	7	0.56	0.0	0.3	0.3	0.3
	True	233	36	7	0.56	0.0	0.3	0.5	0.33
Gaussian	False	233	36	7	0.56	0.0	0.3	0.3	0.3
	True	233	36	7	0.56	0.0	0.3	0.5	0.4

We find that while fairness constraint increases the proportion of women in the finalist (and ultimately the proportion of women that will be hired), the effect is modest and does not reach parity. With no fairness constraint, 30% of hires are female (the same proportion as the applicant pool). With parity constraint, this proportion increases to 0.33 - 0.4, depending on the dependency structure.

6.3.2 ML Agent-Based Estimation

Second, we estimate the effectiveness of the parity constraint using ML agent-based simulation process described in Section 5.4. Unlike in parametric estimation, here, the results are independent of the structural model and all of its assumptions (same quality distribution among men and women, unbiased hiring manager, dependency structure).

We perform the simulation on a subset of jobs where female applicants are underrepresented ($p < 0.5$) using the hold-out test set and report the total resulting gender distribution in Table 7. Overall with no parity constraint, the proportion of female finalists is 0.37. With parity constraint, it increases to 0.42. Note that there is also a high level of heterogeneity in the effectiveness of parity constraint across job types. In Engineering & Technical jobs, for example, the parity constraint only increases the proportion of female finalists to 0.32. We repeat this analysis for a smaller subset of jobs where there is a higher level of female underrepresentation ($p < 0.4$) and report the results in Appendix E.

Table 7: Effectivess of Parity Constraint – Agent-Based Simulation

Job Category	N Jobs	N Apps	N Shortlist	N Finalist	Parity Constraint	Prop. of Females in App. Pool	Shortlist	Finalist
Biz Dev & Operations	14	2,020	322	90	False	0.35	0.4	0.4
					True	0.35	0.5	0.43
Customer Service & Acct Management	16	3,180	402	109	False	0.42	0.51	0.51
					True	0.42	0.5	0.51
Engineering & Technical	97	15,821	3,244	600	False	0.26	0.21	0.24
					True	0.26	0.46	0.32
Finance & Accounting	30	6,262	750	222	False	0.35	0.38	0.4
					True	0.35	0.5	0.48
HR	3	110	34	11	False	0.36	0.47	0.64
					True	0.36	0.5	0.55
Legal & PR	12	1,777	288	57	False	0.27	0.33	0.49
					True	0.27	0.49	0.51
Product & Design	47	9,099	1,594	272	False	0.35	0.34	0.4
					True	0.35	0.5	0.45
Sales & Marketing	34	11,314	1,888	453	False	0.35	0.39	0.45
					True	0.35	0.5	0.48
Overall	253	49,583	8,522	1,814	False	0.32	0.32	0.37
					True	0.32	0.48	0.42

We run the simulation at the job posting level and aggregate it up to the job category level for all jobs with $p < 0.5$ in the hold-out test set. Note that in some categories, under parity constraint, the % of females in the shortlist is not exactly 50% for some jobs since there may be less than $n_s/2$ females in the applicant pool.

6.4 Test of Theoretical Predictions

We test the theoretical predictions of how model parameters affect the probability of a female hire using a simple linear model presented in Section 5.5. The goal is to test if our theoretical predictions still hold with hiring data, where parametric assumptions about the quality distributions of the applicants, dependency structure between screener and hiring manager, etc., do not necessarily hold. We estimate the model with and without parity constraint using logit regression to test the following hypotheses. Note that we estimate the model on the hold-out test set, and in concordance with our theoretical model, we only use the subset of job postings with majority male applicants ($p < 0.5$).

Hypothesis 1. *Parameter θ will have a negative effect on the probability of female hire with parity constraint but no effect without parity constraint.*

Hypothesis 2. *Parameter δ will have a negative effect on the probability of female hire, both with and without parity constraint.*

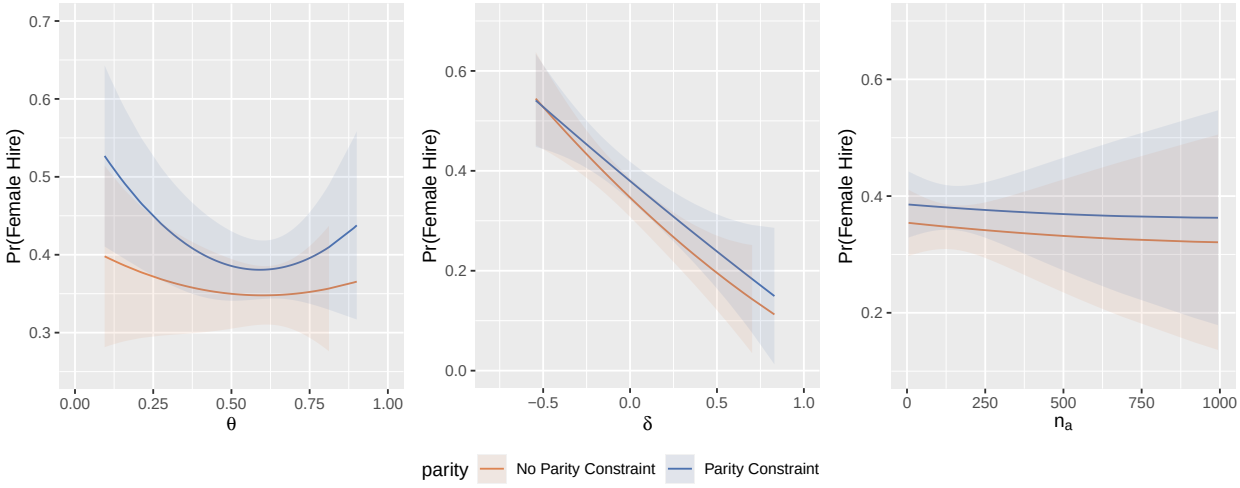
Hypothesis 3. *Parameter n_a will have a positive effect on the probability of female hire with parity constraint but no effect without parity constraint.*

We report the regression estimates in Table 8 and the empirical estimates of predicted probabilities in Figure 15. We find support for Hypothesis 1 since $\hat{\theta}$ is negatively related to the probability of female hire under parity constraint (Column 1) but has no effect when there is no parity constraint (Column 2). We also find support for Hypothesis 2 since δ has a negative relation with the probability of female hire both with and without parity constraint. We do not find empirical support for Hypothesis 3, as we do not find any significant relation between the size of the applicant pool n_a and the probability of female hire. One reason for this is that there may not be enough variation in the size of the applicant pool in the data. Although the theoretical predictions show that the probability of female hire increases with n_a, most of the increase happens initially at lower values of n_a.

Table 8: Probability of Female Hire – Logit Estimates

	Dep Var: $\hat{Pr}$(Female Hire)	
	With Parity Constraint	No Parity Constraint
$\hat{\theta}$	-0.702^{**} (0.279)	-0.235 (0.280)
$\hat{\theta}^2$	0.594^{**} (0.289)	0.196 (0.290)
$\hat{\delta}$	-0.289^{***} (0.064)	-0.332^{***} (0.064)
$\hat{\delta}^2$	0.014 (0.088)	0.060 (0.088)
n_a	-0.00004 (0.0002)	-0.0001 (0.0002)
n_a^2	0.00000 (0.00000)	0.00000 (0.00000)
Controls		
p	1.029^{***} (0.145)	1.370^{***} (0.145)
n_s	0.00001 (0.001)	0.0001 (0.001)
n_f	0.002 (0.002)	0.0003 (0.002)
Constant	0.275^{***} (0.088)	0.007 (0.088)
Observations	254	254
Log Likelihood	-7.296	-7.941
Note:		*p<0.1; **p<0.05; ***p<0.01

Figure 15: Empirical Prediction of Pr(Female Hire)



7 Discussion and Conclusion

We now discuss the managerial and algorithmic design implications of our findings and suggest some paths forward for future work.

First, gender parity in the shortlist does not necessarily lead to gender parity in finalists/hires, even with unbiased hiring managers. So, managers cannot expect to have a gender-balanced workforce by only implementing a parity constraint in the initial stage. Moreover, the lack of gender parity in hires when there is gender parity in the shortlist is not an indication that the hiring manager is biased. Unless this is understood and anticipated by all stakeholders, the parity constraint runs the risk of making unbiased hiring managers *appear* biased since they seem to "undo" the parity constraint in the interview stage.

Second, there is a significant amount of heterogeneity in the effectiveness of parity constraint across jobs, so a blanket parity constraint will have varying effects depending on the job type. Instead, the constraint should be specific to each job or job category.

Third, the effectiveness of the parity constraint *and* the expected quality of hires increases with the size of the applicant pool. This implies that the second-best way to increase the diversity of the workforce (second to increasing the proportion of women in the applicant pool) is to increase the size of the applicant pool. In other words, it's easier for large and selective firms to increase diversity through parity constraints.

Fourth, the effectiveness of the parity constraint decreases as the correlation between the algorithm's and the hiring manager's estimates of candidate quality increases. This means that the better the algorithm becomes at learning the hiring manager's estimate of quality, the less effective the parity constraint becomes. The expected quality of hires also decreases as the correlation increases. A design implication is that the screening algorithm should be trained to learn a score that is as correlated as possible with the candidate's true quality but as orthogonal as possible with the hiring manager's estimate of quality. That is, the screening algorithm's assessment should be *complementary* to the hiring manager's assessment rather than similar.

Finally, imposing a parity constraint on the algorithm does not absolve developers from ensuring that the algorithm is statistically unbiased. Algorithmic fairness constraints are typically taken as a "fix" for algorithmic bias. But, when algorithmic outputs are used as inputs in downstream decisions, algorithmic bias can still negate any parity constraints. As we have shown, the effectiveness of the parity constraint depends on the extent to which the screening algorithm is differentially predictive for female and male candidates, a form of algorithmic bias. Although our empirical results did not show a difference in predictive power, it is still a valid concern especially given that there may not

be as much training data for minority candidates.

Future Work

One of the key insights of this paper is that the screening algorithm should be *complementary* to the hiring manager's assessment rather than similar. A design implication is that the screening algorithm should be trained to learn a screening score that is as correlated as possible with the candidate's true quality but as orthogonal as possible with the hiring manager's estimate of quality. Recent advances in machine learning, such as adversarial learning, have given us the tools and machinery necessary to train such a model (Zhang et al. 2018). The practical challenge is coming up with a good proxy of candidates' true quality (such as job performance metrics) and properly addressing the missing data problem – i.e., job performance data will only be available for candidates that were hired (Cowgill (2020) shows that with noise/variation in the hiring data, the missing data problem can be somewhat mitigated). Future work could study the design and performance of such a screening algorithm.

The complementary nature of the screening algorithm could also cause organizational challenges. Hiring managers may view AI screening algorithms as a means to automate part of their work, in which case they would want the AI to be as similar as possible to their own assessment. Prior work has studied similar tensions in the algorithmic hiring setting (van den Broek et al. 2021). Future work could study Human-AI management when the AI is designed to be explicitly dissimilar but complementary to the human.

Lastly, a defining feature of the model is that the hiring manager is unbiased. This is to show that even when the hiring manager is unbiased, the parity constraint may not be effective. Future work could study whether the introduction of algorithmic fairness constraints *induces* bias in hiring managers who were previously unbiased. The psychology and management literature has documented that in the presence of affirmative action programs (AAPs), the majority groups stigmatize AAP hires and view them as less competent (Heilman et al. 1997; Leslie et al. 2013). These negative views are extended to minorities even if they are not hired under AAPs through stereotyping (Heilman et al. 1997). Since algorithmic fairness constraints can be perceived as variations of AAPs, future work could study whether the introduction of fair algorithms induces bias in hiring managers that were previously unbiased.

Appendix

A Proofs

A.1 Correlation Parameter

Proposition. *The probability that a female is hired decreases when the correlation between screening and hiring manager scores increases.*

Proof. A female is hired when the hiring manager score of the shortlisted female exceeds the shortlisted male's score.

$$\begin{aligned} Pr(\text{Female is hired}) &= Pr(Q^H_{s,f} > Q^H_{s,m}) \\ &= \int_0^1 Pr(Q^H_{s,m} = y) \cdot Pr(Q^H_{s,f} > y)\,\mathrm{d}y \\ &= \int_0^1 f_{Q^H_{s,m}}(y)(1 - F_{Q^H_{s,f}}(y))\,\mathrm{d}y \\ &= \int_0^1 \theta \frac{y^{-1+n_a(1-p)}}{Beta[n_a(1-p),1]} + (1-\theta)(1 - \int \theta \frac{y^{-1+n_a p}}{Beta[n_a p,1]} + (1-\theta))\,\mathrm{d}y\,\mathrm{d}y \end{aligned} \tag{A.1}$$

This gives a closed-form solution:

$$Pr(\text{Female is hired}) = \frac{(2p-1)\theta^2(n_a(p-1)+1)(n_a p-1) + 2n_a(1-2p)\theta + n_a(n_a(p-1)p-1) - 1}{2(n_a(p-1)-1)(n_a p+1)} \tag{A.2}$$

The derivative of the probability of a female hire with respect to θ is negative.

$$\frac{\mathrm{d}Pr(\text{Female is hired})}{\mathrm{d}\theta} = \frac{(2p-1)(n_a(n_a(p-1)p\theta + \theta - 1) - \theta)}{(n_a(p-1)-1)(n_a p+1)} < 0 \tag{A.3}$$

□

A.2 Size of Applicant Pool

Proposition. *When $\theta < 1$, the probability that a female is hired increases with the size of the applicant pool.*

Proof. The derivative of the probability of a female hire with respect to the size of the applicant pool n_a is positive:

$$\frac{\mathrm{d}Pr(\text{Female is hired})}{\mathrm{d}n_a} = -\frac{(2p-1)(\theta-1)\theta\left(n^2(p-1)p+1\right)}{\left(-\left(n^2(p-1)p\right)+n+1\right)^2} > 0 \tag{A.4}$$

□

A.3 Difference in Predictive Power

Proposition. *The probability that a female is hired decreases when the screening algorithm is less predictive for female candidates compared to male candidates.*

Proof. Using different θ for male and female candidate ($\theta_m = \theta$ and $\theta_f = \theta - \delta$), the probability that a female is hired is given by:

$$Pr(\text{Female is hired}) = \frac{\theta(-n(p-1)(\delta-\theta)-1)}{n(p-1)-1} - \frac{(\delta-\theta)(p\theta(n(p-1)+1)-1)}{np+1} + \frac{1}{2}(\theta-1)(\delta-\theta-1) \tag{A.5}$$

The derivate with respect to δ is:

$$\frac{\mathrm{d}Pr(\text{Female is hired})}{\mathrm{d}\delta} = \frac{1}{2}\left(-\frac{2n_a(p-1)\theta}{n_a(p-1)-1} + \frac{2p\theta(n_a(-p)+n_a-1)+2}{n_ap+1} + \theta - 1\right) \tag{A.6}$$

which is negative.

□

A.4 Cost of Parity Constraint

We calculate the expected hiring manager score of hires ($E[Q_h^H]$) with and without parity constraint.

$$\textit{Cost of Parity Constraint} = E[Q_h^H]_{\textit{No Parity Constraint}} - E[Q_h^H]_{\textit{Parity Constraint}} \tag{A.7}$$

First, consider the case where there is a parity constraint in the shortlist.

Parity Constraint

Under parity constraint, the screening algorithm will shortlist the best male candidate and the best female candidate. Given this shortlist, the expected hiring manager score of the hired candidate is:

$$E[Q_h^H] = E[Q_{h,m}^H] \cdot Pr(\textit{Male is hired}) + E[Q_{h,f}^H] \cdot Pr(\textit{Female is hired}) \tag{A.8}$$

We know the probability that a male or female is hired from above. The expected quality of a male hire is:

$$E[Q_{h,m}^H] = \int_0^1 y \frac{f_{Q_m^H}(y) F_{Q_f^H}(y)}{Pr(Q_m^H > Q_f^H)} \, \mathrm{d}y \tag{A.9}$$

$$= \frac{2\left(\frac{3n_a(p-1)\theta(\delta-\theta+1)}{n_a(p-1)-2} + \frac{3(\delta-\theta)(\theta(n_a(p-1)+1)(n_ap+1)-n_a-1)}{(n_a+1)(n_ap+2)} + \theta(-\delta+\theta-1) + \delta - \theta + 1\right)}{3\left(\frac{2(\delta-\theta)(p\theta(n_a(p-1)+1)-1)}{n_ap+1} + \frac{2(n_a(p-1)\theta(\delta-\theta)+\theta)}{n_a(p-1)-1} - \delta\theta + \delta + \theta^2 + 1\right)} \tag{A.10}$$

Similarly, the expected quality of a female hire is:

$$E[Q_{h,f}^H] = \int_0^1 y \frac{f_{Q_f^H}(y) F_{Q_m^H}(y)}{Pr(Q_f^H > Q_m^H)} \, \mathrm{d}y \tag{A.11}$$

$$= \frac{3n_ap(\delta-\theta)\left(\frac{\theta-1}{n_ap+2} - \frac{\theta}{n_a+1}\right) - \frac{3\theta(\delta-\theta+1)}{n_a(p-1)-2} + \theta(-\delta+\theta-1) + \delta - \theta + 1}{3\left(\frac{\theta(-n_a(p-1)(\delta-\theta)-1)}{n_a(p-1)-1} - \frac{(\delta-\theta)(p\theta(n_a(p-1)+1)-1)}{n_ap+1} + \frac{1}{2}(\theta-1)(\delta-\theta-1)\right)} \tag{A.12}$$

No Parity Constraint

Without the parity constraint, the screening algorithm will shortlist the two best candidates. We know the screening score of the best candidate (Q_{s1}^S) has a beta distribution $Q_{s1}^S \sim \text{Beta}[n_a, 1]$. The screening score of the 2nd best candidate (Q_{s2}^S) also has a beta distribution $Q_{s2}^S \sim \text{Beta}[n_a - 1, 2]$.

Combining these, the screening scores of the shortlist has a mixture distribution that is a mix of the two beta distributions.

$$f_{Q_s^S} = 0.5 \cdot f_{Q_{s1}^S} + 0.5 \cdot f_{Q_{s2}^S} \tag{A.13}$$

The corresponding hiring manager score in the shortlist Q_s^H has a distribution that is a mixture of the above and uniform.

$$f_{Q_s^H} = \theta f_{Q_s^S} + (1 - \theta) \tag{A.14}$$

The best candidate from this shortlist gets hired. Therefore, the expected hiring manager score of the hired candidate is:

$$E[Q_h^H] = E[X|X > Y] \tag{A.15}$$

$$X, Y \sim Q_s^H \tag{A.16}$$

$$= \frac{\delta^2 p^2 (n_a(n_a(n_a(1.1875\ - 0.333n_a) - 0.646) - 1) + 0.417)}{(n_a + 1)(n_a + 2)(n_a - 0.5)(n_a + 0.5)} \tag{A.17}$$

$$+ \frac{\delta p (n_a(n_a(n_a(1 - 0.667n_a) + 0.833) - 0.25) - 0.167)}{(n_a + 1)(n_a + 2)(n_a - 0.5)(n_a + 0.5)} \tag{A.18}$$

$$+ \frac{\theta^2 (n_a(n_a(n_a(1.1875 - 0.333n_a) - 0.646) - 1.) + 0.417)}{(n_a + 1)(n_a + 2)(n_a - 0.5)(n_a + 0.5)} \tag{A.19}$$

$$+ \frac{\theta (n_a(n_a(n_a(n_a(0.667\delta p + 0.667) - 2.375\delta p - 1.) + 1.292\delta p - 0.833) + 2\delta p + 0.25) - 0.833\delta p + 0.167)}{(n_a + 1)(n_a + 2)(n_a - 0.5)(n_a + 0.5)} \tag{A.20}$$

$$+ \frac{n_a(n_a((0.667n_a + 2)n_a + 1.167) - 0.5) - 0.333}{(n_a + 1)(n_a + 2)(n_a - 0.5)(n_a + 0.5)} \tag{A.21}$$

B Simulation Algorithms

Algorithm 1: Parametric Simulation

Input: $n_a, n_s, n_h, p, \theta, \delta, parity, dependency$ // model parameters
Output: Pr(Female Hire)
Function ParametricSimulation($n_a, n_s, n_h, p, \theta, \delta, parity, dependency$):

```
/* generate list of applicants and scores                                  */
A ← EmptyList(size=n_a)                    // generate list of applicants of size n_a
A.qs ← Uniform[0,1]                  // generate screening score q^S for each applicant
A[0:n_a p].female ← 1                                  // n_a p applicants are female
A[n_a p:].female ← 0                                 // n_a(1 − p) applicants are male

/* rank order and shortlist applicants by screening score                  */
A ← A.sortBy(qs)
if parity == True then
    /* under parity constraint, shortlist top n_s/2 male and n_s/2 female candidates  */
    Sf ← A[A.female == 1][:n_s/2]                                  // shortlisted males
    Sm ← A[A.female == 0][:n_s/2]                                // shortlisted females
    S← Concatanate(Sm, Sf)                                     // shortlisted canidates
end
else
    /* without parity constraint, shortlist top n_s candidates                 */
    S = A[:n_s]
end

/* hire applicants from the shortlist                                      */
S.qh ← EstimateQh(S.qs, θ, δ, dependency)    // estimate hiring manager score q^H
H ← S.sortBy(qh)[:n_h]                    // hire the top n_h candidates based on q^H
Pr(Female Hire) ← H[H.female == 1].size()/n_h      // proportion of female hires
return Pr(Female Hire)
```

End Function

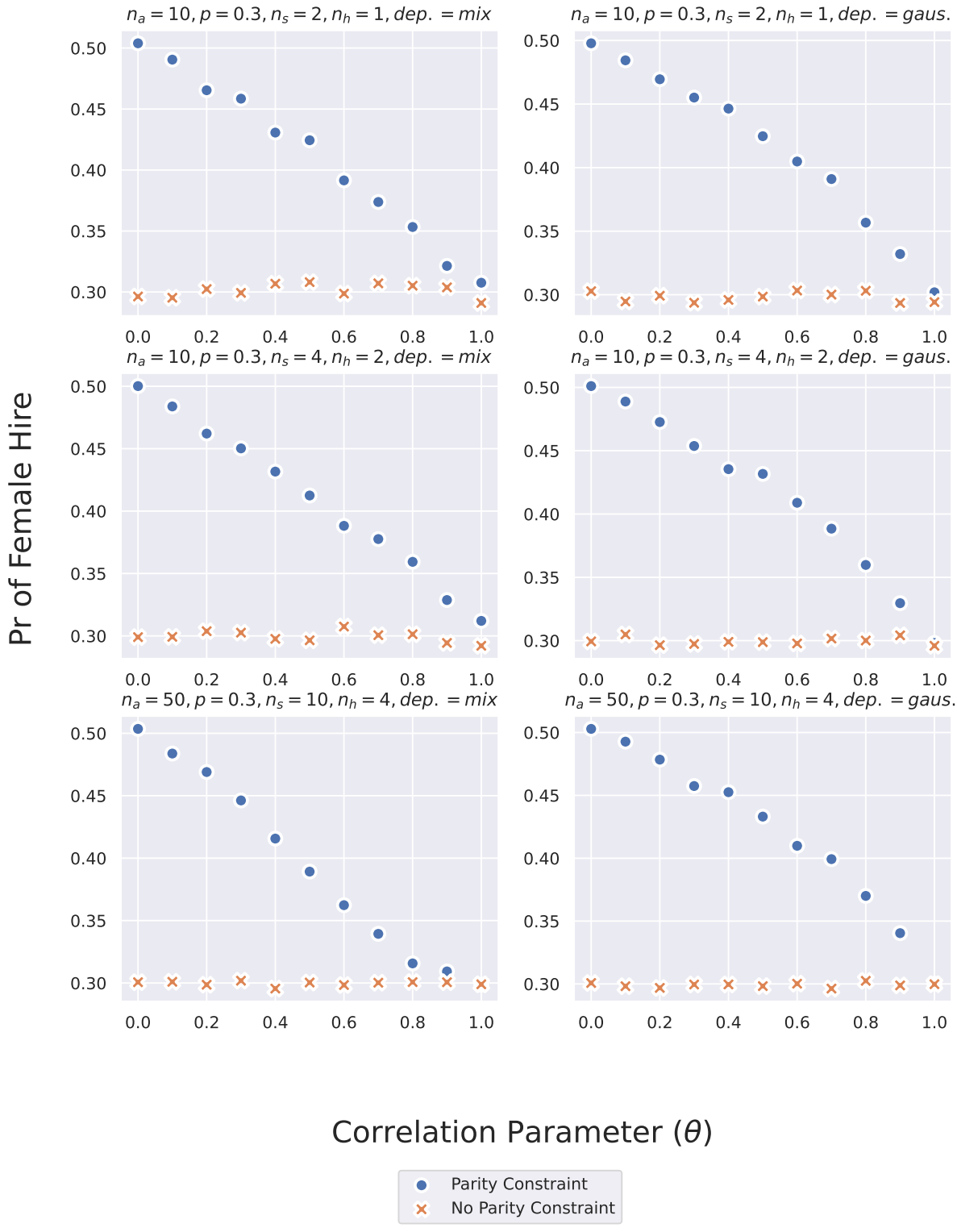
Algorithm 2: ML Agent-Based Simulation

Input: Data, MLScreener, MLHiringManager
Output: `Pr(Female Hire)`
Function `MLAgentBasedSimulation(`*Data, MLScreener, MLHiringManager*`):`

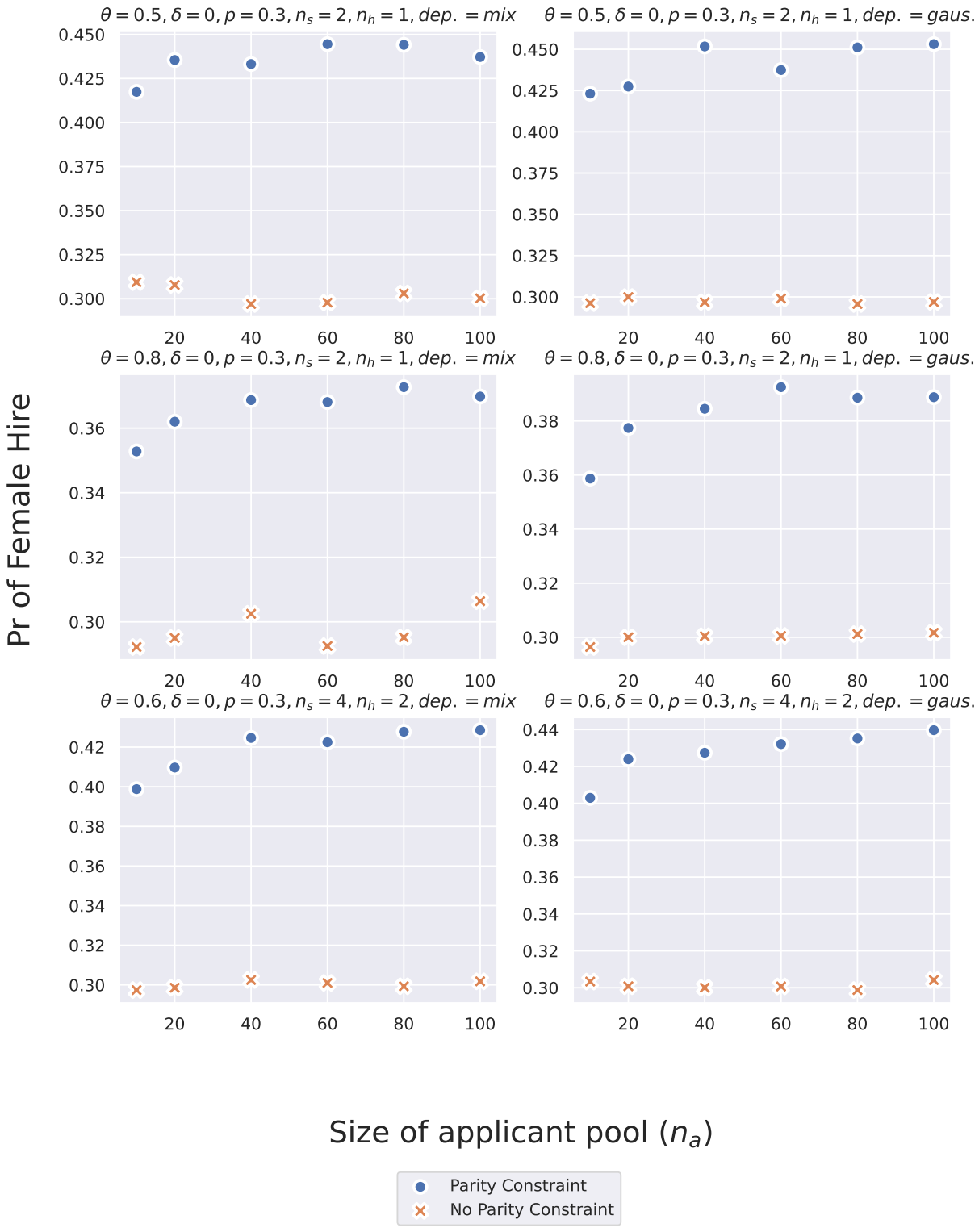
```
    /* get size of applicant pool, shortlist, finalist from data                    */
    n_a ← getNumApps(Data)                                        // applicant pool
    n_s ← getNumShortlist(Data)                                       // shortlist
    if n_s mod 2 ≠ 0 then
        n_s = n_s + 1                          // ensure number of shortlist is even
    end
    n_h ← getNumFinalist(Data)                                         // finalist

    /* get screening scores, gender                                                 */
    A ← EmptyList(size=n_a)                  // generate list of applicants of size n_a
    A.qs ← MLScreener.estimateQs(Data)                  // predict screening scores
    A.female ← getGenders(Data)                                      // get genders

    /* rank order and shortlist applicants by screening score                       */
    A ← A.sortBy(qs)
    if parity == True then
        /* under parity constraint, shortlist top n_s/2 male and n_s/2 female candidates  */
        Sf ← A[A.female == 1][:n_s/2]                              // shortlisted males
        Sm ← A[A.female == 0][:n_s/2]                            // shortlisted females
        S← Concatanate(Sm, Sf)                                 // shortlisted canidates
    end
    else
        /* without parity constraint, shortlist top n_s candidates                  */
        S = A[:n_s]
    end

    /* interview applicants from the shortlist                                      */
    S.qh ← MLHiringManager.estimateQh(Data)      // predict hiring manager scores q^H
    H ← S.sortBy(qh)[:n_h]                 // interview the top n_h candidates based on q^H
    Pr(Female Hire) ← H[H.female == 1].size()/n_h      // proportion of female hires
    return Pr(Female Hire)
End Function
```

C Parametric Simulation Results

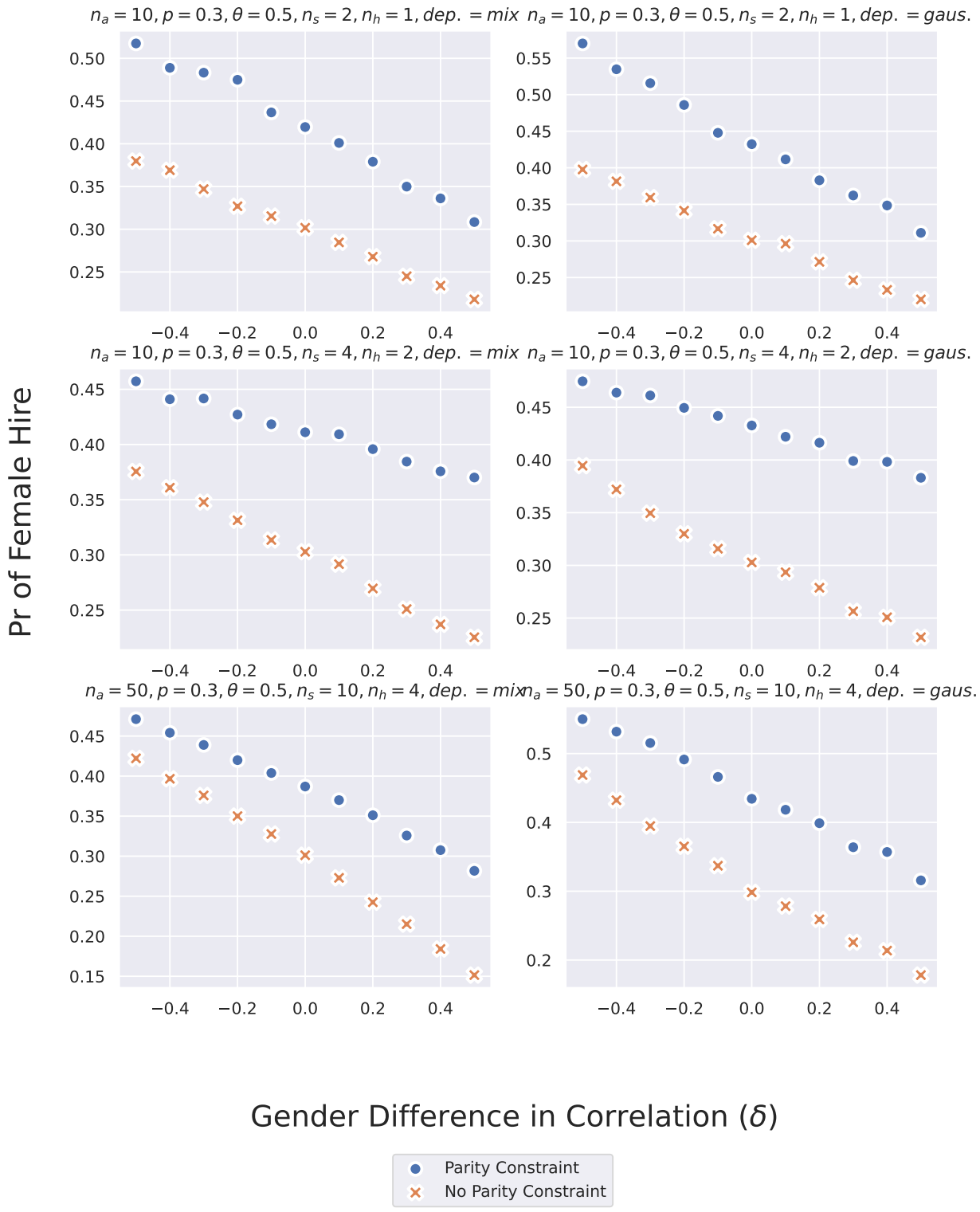
Probability of Female Hire vs. Correlation θ



Probability of Female Hire vs. Size of applicant pool n_a



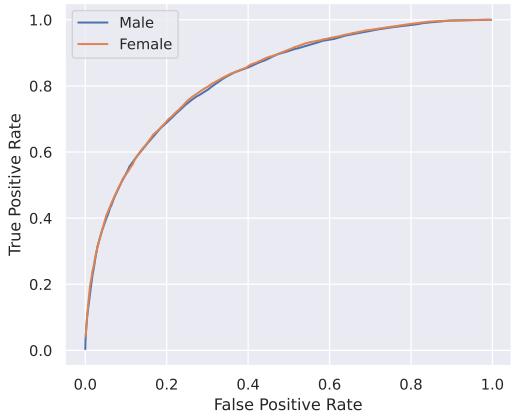
Probability of Female Hire vs. Gender Difference in Predictive Power δ



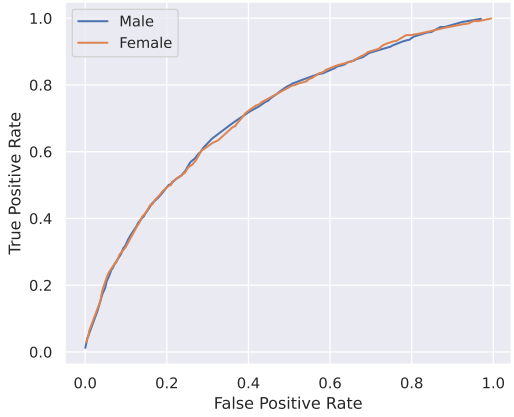
D Screening and Hiring Manager ML Models

D.1 ROC Curves

Screening Model ROC Curve



Hiring Manager Model ROC Curve



D.2 Predictive Features using SHAP Values

To qualitatively assess the models, we recover the most predictive features from the screening and hiring manager models using an explainability method called SHAP (Shapley Additive Explanations) (Lundberg and S.-I. Lee 2017).

SHAP uses a "perturbation" approach to estimate how a perturbation in the feature space changes the prediction. For example, if we remove a particular token from a resume (e.g. "stanford")

and it drastically changes the predicted probability of being shortlisted, then the token receives a high SHAP value for that particular prediction. In a different resume, removing the same token may not change the predicted probability at all, in which case, the same token receives a SHAP value closer to zero. Although SHAP values provide explainability at the instance level (i.e. for a given token in a given resume), we can aggregate across instances (resumes) to get to model-level explainability.

Table 9: Most Predictive Features

Job Correlation	Customer Experience Lead $\theta = 0.22$		Entry-level Software Engineering $\theta = 0.51$		Software Internship $\theta = 0.76$	
	Screening	Hiring Manager	Screening	Hiring Manager	Screening	Hiring Manager
1	medicare	fraud	ann arbor	san francisco	square	san francisco
2	d	referral	seattle	led	paypal	stanford
3	medicaid	led	added	pittsburgh	cambridge	fraud
4	aetna	advocate	inc	openstack	sunnyvale	oversaw
5	fraud	across	internal	docker	freshmen	fundraising
6	present	resolved	mi	uber	princeton	stripe
7	claims	inbound	amazon	storm	seattle	university of california, berkeley
8	providers	child	june 2017	salesforce	inc	freshmen
9	appeals	volunteer	present	jenkins	ebay	salesforce
10	drug	lead	honor	pipeline	apple	berkeley
11	chat	teams	metadata	lambda	cornell	uc berkeley
12	eligibility	outbound	acm	overflow	waterloo	volunteers
13	internal	liaison	references	seattle	kappa	founded
14	prescription	aetna	scholar	redis	ann arbor	attendees
15	medication	third	advisor	deans	phi	founder
16	referral	authorization	wa	lead	2017-2018	raised
17	authorizations	back	austin	four	interns	transition
18	inbound	party	scalable	ann arbor	stanford	teammates
19	therapy	terms	08/2016	across	providence	freshman
20	enrollment	authorizations	docker	legacy	october 2016	outreach

We report the most predictive features for three jobs at varying levels of correlation in Table 9 (We describe how we estimate these correlations for each job in the next section). The first job is Customer Experience Lead, which has a low correlation between screening and interview ($\hat{\theta} = 0.22$). The screening criteria for this job appear to be the candidate's past experience in the insurance-related domain (e.g., `medicare, (part) d, medicare, aetna, claims`). In the interview stage, the criteria appear to be functional skills (i.e., what candidates did) as well as leadership skills (e.g., `led, lead, advocate, resolved`).

The second job is an entry-level software engineering job with a medium level of correlation ($\hat{\theta} = 0.51$). The screening criteria for this job appears to be location (`ann arbor, seattle, austin`), educational achievements (`scholar, acm`), and past jobs/internships (`amazon`). The interview criteria includes location (`san francisco, seattle, ann arbor`), past jobs (`uber, salesforce`),

but also technical skills (`openstack, docker, redis, storm`).

The third job is a software engineering internship with a high level of correlation ($\hat{\theta} = 0.76$). For both screening and interview, the candidate's university and past internships are highly predictive features.

E Simulation Results

Table 10: Parametric Simulation by Job Category ($p < 0.5$)

Job Category	Prop. of Females in App Pool	Shortlist	Finalist
Biz Dev & Operations	0.37	0.5	0.42
Customer Service & Acct Management	0.38	0.5	0.45
Engineering & Technical	0.23	0.5	0.34
Finance & Accounting	0.35	0.5	0.43
HR	0.35	0.5	0.45
Legal & PR	0.32	0.5	0.41
Product & Design	0.33	0.5	0.41
Sales & Marketing	0.36	0.5	0.42

Table 11: Effectivess of Parity Constraint – Agent-Based Simulation ($p < 0.4$)

Job Category	N Jobs	N Apps	N Shortlist	N Finalist	Parity Constraint	Prop. of Females in App. Pool	Shortlist	Finalist
Biz Dev & Operations	7	1,042	162	51	False	0.26	0.35	0.35
					True	0.26	0.5	0.41
Customer Service & Acct Management	8	983	130	53	False	0.32	0.42	0.45
					True	0.32	0.5	0.43
Engineering & Technical	93	14,515	3,030	566	False	0.24	0.2	0.22
					True	0.24	0.45	0.31
Finance & Accounting	21	4,627	568	173	False	0.32	0.34	0.34
					True	0.32	0.5	0.44
HR	2	52	16	6	False	0.29	0.38	0.33
					True	0.29	0.5	0.33
Legal & PR	8	1,397	212	33	False	0.22	0.24	0.36
					True	0.22	0.49	0.39
Product & Design	32	5,480	1,084	182	False	0.28	0.28	0.32
					True	0.28	0.49	0.41
Sales & Marketing	21	8,656	1,230	297	False	0.32	0.34	0.41
					True	0.32	0.5	0.46
Overall	192	36,752	6,432	1,361	False	0.28	0.26	0.31
					True	0.28	0.48	0.38

Table 12: Effectivess of Parity Constraint – Agent-Based Simulation (Adjusted Shortlist)

Job Category	N Jobs	N Apps	N Shortlist	N Finalist	Parity Constraint	Applicant Pool	Shortlist	Finalist
Biz Dev & Operations	14	2,020	322	90	False	0.35	0.4	0.4
					True	0.35	0.5	0.43
Cust Serv & Acct Mgmt	16	3,180	402	109	False	0.42	0.51	0.51
					True	0.42	0.5	0.51
Engineering & Technical	97	15,821	2,964	600	False	0.26	0.22	0.23
					True	0.26	0.5	0.37
Finance & Accounting	30	6,262	748	222	False	0.35	0.38	0.4
					True	0.35	0.5	0.48
HR	3	110	34	11	False	0.36	0.47	0.64
					True	0.36	0.5	0.55
Legal & PR	12	1,777	284	57	False	0.27	0.34	0.49
					True	0.27	0.5	0.51
Product & Design	47	9,099	1,580	272	False	0.35	0.34	0.4
					True	0.35	0.5	0.45
Sales & Marketing	34	11,314	1,886	453	False	0.35	0.39	0.45
					True	0.35	0.5	0.48
Overall	253	49,583	8,220	1,814	False	0.32	0.32	0.37
					True	0.32	0.5	0.44

References

Angwin, Julia and Jeff Larson (2022). "Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say *". In: *Ethics of Data and Analytics*. Auerbach Publications.

Blum, Avrim, Kevin Stangl, and Ali Vakilian (June 20, 2022). "Multi Stage Screening: Enforcing Fairness and Maximizing Efficiency in a Pre-Existing Pipeline". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. New York, NY, USA: Association for Computing Machinery, pp. 1178–1193.

Celis, L. Elisa et al. (Mar. 3, 2021). "The Effect of the Rooney Rule on Implicit Bias in the Long Term". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, pp. 678–689.

Chouldechova, Alexandra (June 1, 2017). "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments". In: *Big Data* 5.2, pp. 153–163.

Cowgill, Bo (2020). "Bias and Productivity in Humans and Algorithms: Theory and Evidence from Re´sume´ Screening".

Dastin, Jeffrey (Oct. 10, 2018). "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women". In: *Reuters*.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta (Mar. 16, 2015). "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination". arXiv: 1408.6491 [cs].

Devlin, Jacob et al. (May 24, 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv: 1810.04805 [cs].

Dwork, Cynthia et al. (Jan. 8, 2012). "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. New York, NY, USA: Association for Computing Machinery, pp. 214–226.

Fershtman, Daniel and Alessandro Pavan (Mar. 1, 2021). ""Soft" Affirmative Action and Minority Recruitment". In: *American Economic Review: Insights* 3.1, pp. 1–18.

Geyik, Sahin Cem, Stuart Ambler, and Krishnaram Kenthapadi (Apr. 30, 2019). "Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search". arXiv: 1905.01989 [cs].

Hardt, Moritz, Eric Price, and Nathan Srebro (Oct. 7, 2016). "Equality of Opportunity in Supervised Learning". arXiv: 1610.02413 [cs].

Heilman, Madeline E., Caryn J. Block, and Peter Stathatos (June 1, 1997). "The Affirmative Action Stigma Of Incompetence: Effects Of Performance Information Ambiguity". In: *Academy of Management Journal* 40.3, pp. 603–625.

Huet, Ellen (Jan. 10, 2017). "Facebook's Hiring Process Hinders Its Effort to Create a Diverse Workforce". In.

Joe, Harry (June 26, 2014). *Dependence Modeling with Copulas*. CRC Press. 483 pp. Google Books: 09ThAwAAQBAJ.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (Nov. 17, 2016). "Inherent Trade-Offs in the Fair Determination of Risk Scores". arXiv: 1609.05807 [cs, stat].

Kleinberg, Jon and Manish Raghavan (Jan. 4, 2018). "Selection Problems in the Presence of Implicit Bias". arXiv: 1801.03533 [cs, stat].

Lambrecht, Anja and Catherine Tucker (July 2019). "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads". In: *Management Science* 65.7, pp. 2966–2981.

Lee, Logan M. and Glen R. Waddell (Apr. 1, 2021). "Diversity and the Timing of Preference in Hiring Decisions". In: *Journal of Economic Behavior & Organization* 184, pp. 432–459.

Leslie, Lisa M., David M. Mayer, and David A. Kravitz (July 23, 2013). "The Stigma of Affirmative Action: A Stereotyping-Based Theory and Meta-Analytic Test of the Consequences for Performance". In: *Academy of Management Journal* 57.4, pp. 964–989.

Lundberg, Scott and Su-In Lee (Nov. 24, 2017). "A Unified Approach to Interpreting Model Predictions". arXiv: 1705.07874 [cs, stat].

Mitchell, Shira et al. (2021). "Algorithmic Fairness: Choices, Assumptions, and Definitions". In: *Annual Review of Statistics and Its Application* 8.1, pp. 141–163.

Nelsen, Roger B. (June 10, 2007). *An Introduction to Copulas*. Springer Science & Business Media. 277 pp. Google Books: yexFAAAAQBAJ.

Peng, Andi et al. (Oct. 28, 2019). "What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (1), pp. 125–134.

Raghavan, Manish et al. (Jan. 27, 2020). "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. New York, NY, USA: Association for Computing Machinery, pp. 469–481.

Schuck, Peter H. (2002). "Affirmative Action: Past, Present, and Future". In: *Yale Law & Policy Review* 20.1, pp. 1–96.

Shi, Wei et al. (2018). "The Adoption of Chief Diversity Officers among S&P 500 Firms: Institutional, Resource Dependence, and Upper Echelons Accounts". In: *Human Resource Management* 57.1, pp. 83–96.

Sühr, Tom, Sophie Hilgard, and Himabindu Lakkaraju (July 21, 2021). "Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. New York, NY, USA: Association for Computing Machinery, pp. 989–999.

Sun, Chi et al. (2019). "How to Fine-Tune BERT for Text Classification?" In: *Chinese Computational Linguistics*. Ed. by Maosong Sun et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 194–206.

Teodorescu, Mike et al. (Sept. 1, 2021). "Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation". In: *MIS Quarterly* 45, pp. 1483–1500.

Van den Broek, Elmira, Anastasia Sergeeva, and Marleen Huysman (Sept. 2021). "When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring". In: *MIS Quarterly* 45.3, pp. 1557–1580.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., pp. 5998–6008.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, et al. (Apr. 3, 2017). "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW

'17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, pp. 1171–1180.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, et al. (Nov. 28, 2017). "From Parity to Preference-based Notions of Fairness in Classification". arXiv: 1707.00010 [cs, stat].

Zaheer, Manzil et al. (2020). "Big Bird: Transformers for Longer Sequences". In: Neural Information Processing Systems (NeurIPS).

Zemel, Rich et al. (May 26, 2013). "Learning Fair Representations". In: *International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 325–333.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell (Dec. 27, 2018). "Mitigating Unwanted Biases with Adversarial Learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New York, NY, USA: Association for Computing Machinery, pp. 335–340.

Zheng, Stephan et al. (Apr. 28, 2020). *The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies*. URL: https://arxiv.org/abs/2004.13332v1 (visited on 08/29/2022). preprint.