Hiring Algorithms in Practice: Bias and Homogeneity

Abstract

Many employers use algorithms, often built by the same few algorithm vendors, to screen job applicants. We acquire and analyze a novel dataset of 3,372,132 applicants submitting 4,197,168 applications, where all the applications were mediated by algorithms built by the same vendor. We conjecture that this algorithmic monoculture could lead to the same individuals and groups being consistently overlooked. We find clear racial disparities in applicant outcomes: of all applications submitted by Black applicants, 26.92% are submitted to positions that are significantly biased against Black candidates. Further, individuals also receive homogeneous outcomes: 4% of applicants that apply to 10 positions are rejected from every position they apply to. Based on a large-scale simulation of the the counterfactual outcomes applicants would have received if they applied to many more positions than they did in practice, we find that all applicants would need to apply to 55 positions to ensure they are recommended for at least one position. As the first large-scale study of algorithmic hiring data, our work establishes significant empirical evidence to warrant increased transparency into pervasive and high-stakes hiring algorithms.

1 Introduction

Over 90% of U.S. employers rely on hiring algorithms to screen or rank job applicants [Fuller et al., 2021]. Hiring algorithms shape which applicants are considered for an interview and which are never seen by a human [Autor and Scarborough, 2008, Hoffman et al., 2018]. They function as a bottleneck to employment opportunity for hundreds of millions of workers around the world [Jain et al., 2024].

Given the concentrated hiring algorithm market, many employers procure hiring algorithms from the same few third-party vendors. For example, over 60% of the Fortune 100 and eight of the ten largest US federal agencies use HireVue's algorithms [Nawrat, 2023]. Prior work defines *algorithmic monoculture* as the condition where different employers rely on the same, or very similar, hiring algorithms [Kleinberg and Raghavan, 2021, Bommasani et al., 2022]. Candidates applying to multiple jobs may (possibly unknowingly) be subject to this monoculture by encountering the same vendor's algorithms repeatedly and, therefore, the same biases and quirks, which may screen them out from employment [Creel and Hellman, 2022, Bommasani et al., 2022].

To better understand algorithmic hiring in practice, we conducted a large-scale empirical analysis of algorithmic hiring using a novel dataset acquired from the talent acquisition platform pymetrics. The dataset records 4,197,168 job applications submitted by 3,372,132 applicants to 1,746 positions. Each application is assessed by a machine learning model that assigns a probabilistic score from 0 to 1, which is thresholded at 0.5 into outcomes of "recommend" or "do not recommend". pymetrics provides these recommendations to their clients, informing their decisions on whether to interview and/or hire applicants: based on our data, the recommendations influence hiring decisions at 156 employers spanning 11 market sectors with a total annual revenue of $225 billion dollars.

We find clear evidence of racial bias in the algorithmic recommendations. Prior work has found slight racial bias in pymetrics algorithms [Wilson et al., 2021, Kassir et al., 2023], but not at levels that would concern US employment law. In contrast to these works, which report results for pymetrics as a whole, we disaggregate our analysis on a per-position basis: since pymetrics mediates the decisions

Preprint. Under review.

of 156 distinct employers for 1,746 distinct positions, we report results for each position individually. We show that 17.62% positions demonstrate significant bias against Black applicants: 31.83% of Black applicants apply to at least one position that is significantly biased against applicants.

These findings align with prior evidence of racial bias in hiring decisions. However, in contrast to well-known work that studies decisions mediated by applicant resumes that embed demographic factors (e.g. racialized names like Jamal, stereotyped activities like softball) , pymetrics has applicants play online games that do not overtly embed demography. These results motivate further interrogation of whether algorithmic hiring leads to fairer outcomes given that companies and researchers have posited that algorithms are more objective .

Building on these findings, we empirically study homogeneous outcomes, which prior work has theorized as a byproduct of hiring algorithms [Bommasani et al., 2022]. Due to the algorithmic monoculture in the hiring algorithm supply chain, where many employers depend on the same few algorithm vendors (e.g.pymetrics, HireVue), applicants may receive the same outcome when they apply to different employers due to hiring decision being based on the same or similar hiring algorithms. We provide the first evidence of homogeneous outcomes at scale in deployed hiring algorithms. For example, of applicants that apply to six positions, 5% are rejected from all positions. Further, we establish that as we consider applicants who apply to more and more positions, the rate at which applicants are not recommended for any position falls exponentially (). Our quantitative results reinforce prior theoretical work [Ajunwa, 2021, Bommasani et al., 2022]: since being rejected from every position one applies to is likely to yield extended unemployment, future work should urgently explore the implications of the economy-wide adoption of hiring algorithms in reconstituting overall labor market outcomes.

While our work reveals significant evidence for concern about deployed hiring algorithms, these algorithms enable new opportunities to understand hiring in sophisticated ways. To explore these possibilities, we introduce a new method for studying hiring by simulating counterfactual applicant outcomes. Since applicants generally apply to sparse subsets of all positions (e.g. due to geographic constraints or occupational preferences), we simulate their outcomes as if they had applied to all pymetrics positions. This approach capitalizes on the centralized pymetrics data, the deterministic replicability of algorithmic recommendations, and the efficiency of algorithmic decision-making: while such an approach would be prohibitively costly using traditional social science methods, we are now able to understand what would have been if applicants applied in different ways. Based on our simulation, if applicants applied to literally every pymetrics position, then no applicant would be systemically rejected. Since this entails applicants applying to a very number of positions, significantly beyond any observed behavior, we refine our simulation to have applicants apply to more positions but not every possible position. We find that applicants need to apply to at least 55 positions mediated by pymetrics to guarantee at least one pymetrics recommendation and pass this initial screening stage in the hiring process.

Our work sheds light on hiring algorithms in practice, in turn substantiating the value of independent scrutiny and research. Patterns of bias and systemic rejection, which may be precipitated by hiring algorithms, warrant attention given the established harms of unemployment [Jacobson et al., 1993, Sen, 1997, Sullivan and Von Wachter, 2009]. As policymakers govern hiring algorithms (e.g. New York City Local Law 144 of 2021, Illinois HB 3773 of 2024 that amends the Illinois Human Rights Act, the EU AI Act of 2024), we recommend they mandate researcher access into hiring algorithms: increased transparency into algorithmic hiring is vital for understanding how AI changes employment and identifying new sources of discrimination.

2 Data

To study algorithmic hiring, we analyze four years of data from the hiring algorithm vendor pymetrics. Our data involves three entities: (i) job-seeking applicants, (ii) job-providing employers, and (iii) the hiring algorithm vendor pymetrics. Employers that are clients of pymetrics direct their applicants to the pymetrics platform, where the applicant plays a series of assessment games. pymetrics builds hiring algorithms to make recommendations based on how applicants play the assessment games. Employers use pymetrics's recommendations to make hiring decisions.

Employers. Employers hire pymetrics to build a machine learning model that screens applications with either "recommend" or "do not recommend" recommendations. These recommendations inform the employer's decisions to advance or reject applicants. Our data tracks 156 employers that hire for 11.2 positions on average with each position receiving 1932 applications on average.

Applicants. Applicants applying to pymetrics-mediated positions play either 12 or 16 assessment games that generate high-dimensional gameplay features about the applicant. These features are stored such that if the applicant applies to another pymetrics-mediated position within the next 330 days, the same features will be used. Our data tracks 3,372,132 applicants who submit 1.24 applications on average.

pymetrics. pymetrics builds 16 online games to measure cognitive traits of applicants. For each client, pymetrics trains a binary classifier: positive training examples correspond to the gameplay features of 50 current employees in that role and negative training examples correspond to the gameplay features of random profiles in the pymetrics database.

Recommendations. When an applicant applies to a position, pymetrics runs the associated model on the applicant's gameplay features, yielding a probabilistic score $p \in [0, 1]$. The score p is thresholded at $t = 0.5$: scores above 0.5 yield predictions $y = 1$ ("recommend") and scores below 0.5 yield predictions $y = 0$ ("do not recommend"). On average, 50% of applicants are "not recommended" at this step, meaning that a hiring manager is unlikely to look at their application.

Sample. Overall, our data tracks 4,197,168 applications with the applicant's features, metadata about the applicant, the position name and employer, metadata about the position and employer, the score, and the binary recommendation, which we detail in the supplement. Most relevant to our work, 40.2% of applicants self-report race with a breakdown of 16.8% Asian, 14.2% White, 3.6% Black, 3.0% Hispanic, and all other racial categories below 2% (i.e. fewer than 100,000 applicants). In addition, to understand applicant outcomes across all of their applications, we leverage shared applicant-level identifiers that link applicants when applying to distinct positions and employers.

3 Results

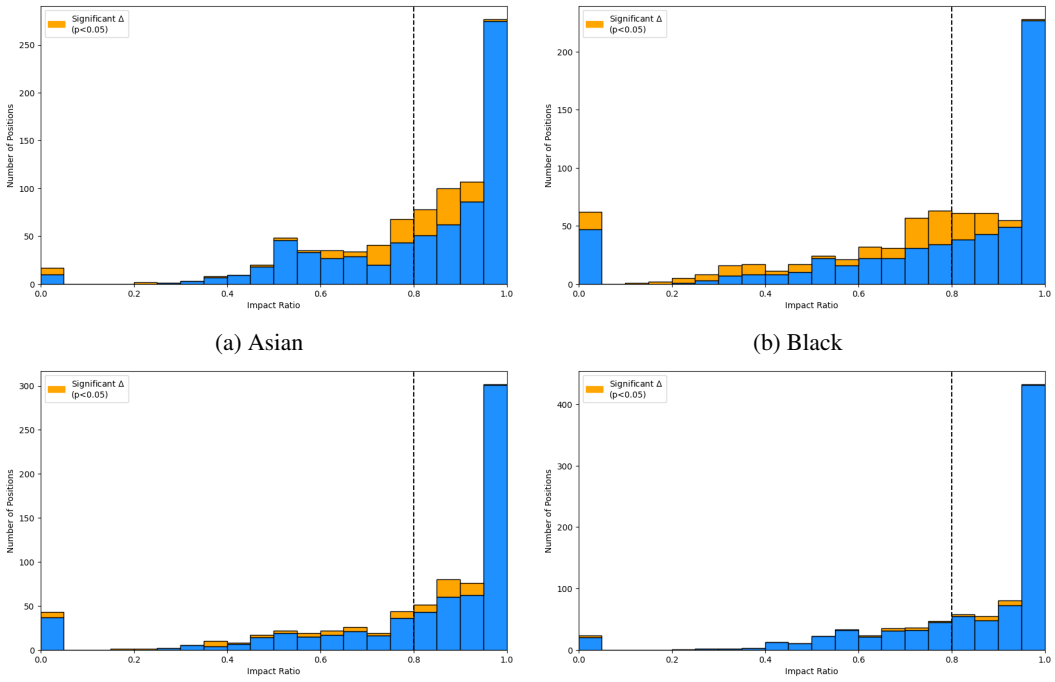
We study the data using three lens: (i) standard analyses in hiring, (ii) unique considerations for algorithmic hiring, and (iii) unique possibilities enabled by algorithmic hiring.

Racial bias. Hiring practices often reflect societal biases that yield worse outcomes for marginalized groups. These biases have been rigorously established and are the subject of anti-discrimination law. .

To demonstrate bias, U.S. legal convention relies on showing that the disadvantaged group is selected significantly less often than the most selected group. As a concrete quantitative standard for bias, we use the well-known guidance of the U.S. Equal Employment Opportunity Commission (EEOC). We define the *selection rate* as the fraction of candidates that are recommended, meaning those whose probabilistic scores exceeds 0.5, based on pymetrics practices. We say a demographic group g is subject to *adverse impact* if: (i) the selection rate for g is less than four-fifths that of the most selected group g' **and** (ii) there is a statistically significant difference at $\alpha = 0.05$ between the selection rates for g and g' according to the two-sample pooled-proportion z-test.

Analyzing the pymetrics data as whole, we find that selection rates are as follows: Asian (53.30%), Black (52.50%), Hispanic/Latino (56.80%), White (58.30%). The rates are lower for marginalized groups (namely for Blacks and Asians), but the disparities do not fall below the EEOC standard since the impact ratios are at least 0.8 (Black impact ratio: 0.839; Asian impact ratio: 0.870). These results align with prior analyses of pymetrics data published by pymetrics [Kassir et al., 2023], researchers [Wilson et al., 2021], and independent auditors [].

However, we argue that providing statistics only for the aggregated data as a whole, while common in prior work, is incomplete and not the appropriate primary lens for understanding hiring algorithm vendors like pymetrics. As an illustrative example, imagine that the recommendations showed a male bias for doctor positions and a female bias for nurse positions. Even if these biases "cancel out" such that the all-position selection rates for men and women are similar, the per-position biases warrant



(a) Asian (b) Black

(c) Hispanic/Latino (d) White

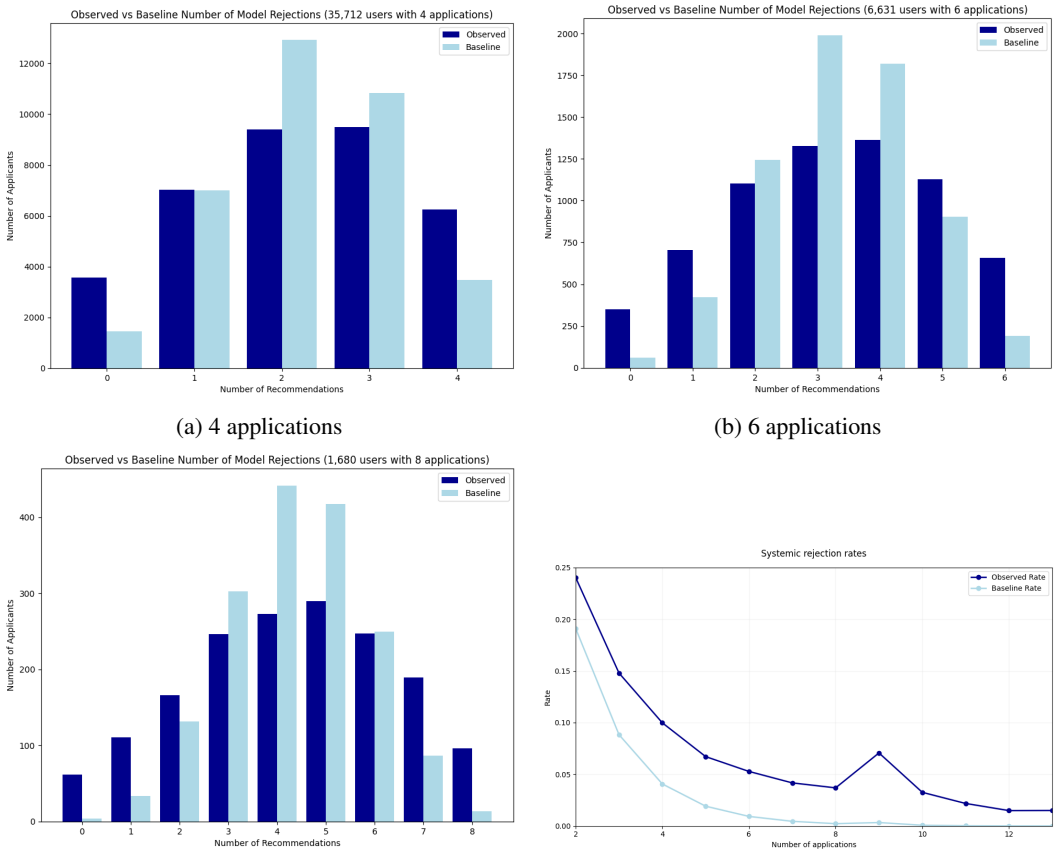
Figure 1: **Racial bias in selection rates.** The histogram for each racial group corresponding to the impact ratio for the group for each of the disaggregated positions. Emphasis is placed on impact ratios below 0.8 and where the group's selection rate is significantly different from that of the most-selected group for $p < 0.05$.

concern. Therefore, since pymetrics is not a singular employer, but instead mediates the decisions of many employers for many positions, we should primarily analyze the pymetrics recommendations for each position *separately* rather than as a single monolithic dataset.

Finding 1: Disaggregation reveals significant racial bias, especially against Black applicants. In Figure 1, we report the histogram of impact ratios for each position mediated by pymetrics across different demographic groups. By disaggregating the pymetrics data by position, we surface clear differences in the outcomes for demographic groups. These differences indicate the largest adverse impact for Black applicants on a relative basis and Asian applicants on an absolute basis.[1] Of all positions, we find adverse impact in 17.62% positions against Black applicants: 31.83% of Black applicants apply to at least one such position. Overall, Black applicants are submitting an application to a position with adverse impact against their racial group 26.92% of the time (41,611 applications). We see similar effects for Asians: 8.49% of all positions demonstrate adverse impact against Asians, affecting 20.55% of Asian applicants and 16.28% of Asian applications (127,312 applications). If applicants to the biased models were counterfactually recommended at the same rate as the most selected racial group, then an additional 12,080 Black applications and 32,357 Asian applications would have been recommended.

Our findings for racial bias indicate the opposite conclusion from prior work []. While prior work generally finds some bias in pymetrics recommendations, it often concludes the bias is fairly marginal and does not rise to concerning levels when compared to the EEOC standard. In contrast, disaggregation surfaces a significant fraction of the recommendations warrant concern given the EEOC standard. The bias in selection is especially striking since pymetrics does not use demographic information as a feature to their model. While prior work in hiring has established that inputs like resumes encode demographic proxies [Bertrand and Mullainathan, 2004], our work provides the first evidence that gameplay features are not only unevenly distributed across racial groups, but that the resulting recommendations yield considerably different selection rates across races. These findings reinforce

[1] More Asians apply than Blacks to pymetrics-mediated positions, hence the discrepancy between relative and absolute.



(a) 4 applications

(b) 6 applications

(c) 8 applications

(d) Systemic rejection rates

Figure 2: **Homogeneous outcomes.**

the broader literature on bias in AI, which has repeatedly shown bias even where demography is not explicitly or obviously encoded into input features [Caliskan et al., 2017, Hofmann et al., 2024].

Homogeneous outcomes. Job seekers generally apply to multiple positions [Dalton and Groen, 2020]. What happens when their job applications are all mediated by hiring algorithms from the same hiring algorithm vendor? Prior work conjectures that such *algorithmic monoculture* [Kleinberg and Raghavan, 2021] is likely to yield *homogeneous outcomes* [Bommasani et al., 2022] where some candidates are recommended to every job they apply to and others are not recommended for any job they apply to. The latter case is very concerning because not being recommended by the first stage screening algorithm is very likely to result in the candidate being rejected by the employer [Fuller et al., 2021]. Therefore, an applicant that is exclusively not recommended by pymetrics algorithms is "algorithmically blackballed" [Ajunwa, 2021] and, quite possibly, locked out of finding a new job.

Therefore, systemic rejection across all of an applicant's applications could contribute to extended employment, which has been shown to be seriously harmful to individuals in a broad, multidisciplinary literature [Kroft et al., 2013a, Sen, 1997]. . Those who are unemployed for extended periods experience discouragement, depleted financial resources, and deteriorating physical and mental health [Rothstein, 2016]. They are also more likely to exit the labor force. Beyond the harms to individuals, if systemically rejected workers remain out of work for long periods and are unable to contribute to the labor force, this may pose broader economy-wide harms to GDP due to the misallocation of talent [Hsieh et al., 2019].

Finding 2: Homogeneous outcomes are frequent. Empirically, we observe a significant rate of homogeneous outcomes and systemic rejections in the pymetrics data by studying the overall outcomes for applicants who apply to multiple positions (Figure 2). For example, of all applicants to apply to 4 positions mediated by pymetrics, 10% are not recommended for any position and 17% are recommended for every position. To better reason about these observed rates, we introduce a *baseline*

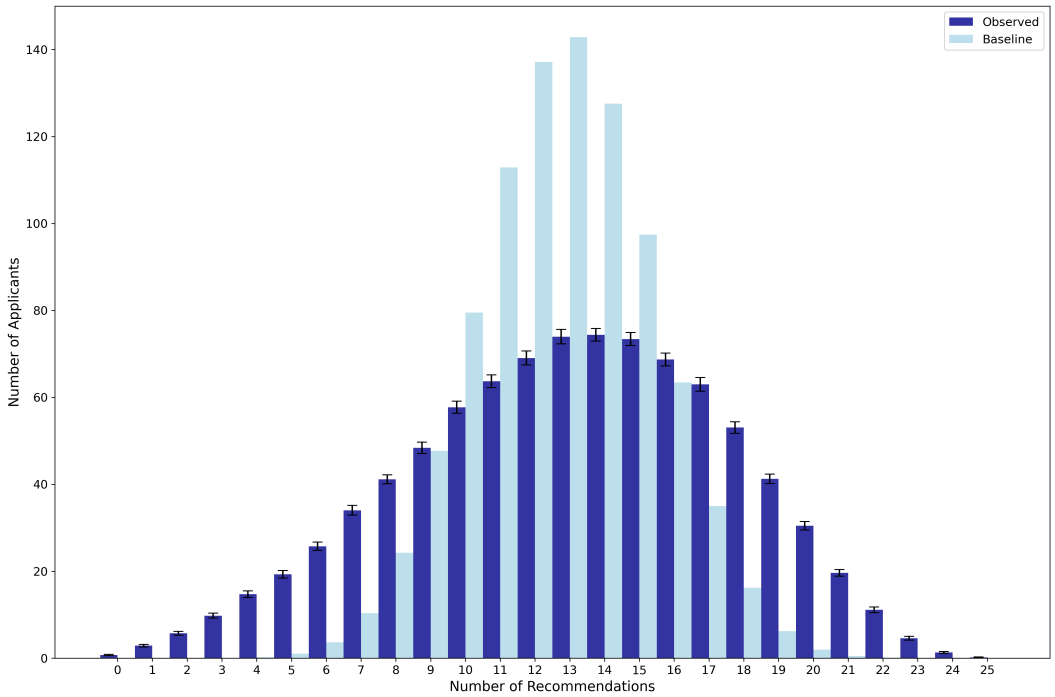


Figure 3: **Simulated outcomes for applicants.** The simulated outcomes for a random subset of applicants if they were evaluated by all connected models to the original set of models they applied to, where two models are connected if at least one person applied to positions associated with both models. For clarity, we only consider applicants whose connected set contains at least 25 models and, if it exceeds 25, we consider their outcomes for a random subset of 25 models with associated 95% confidence interval.

reference. In comparison to the rate at which the same applicant is rejected from all k positions they apply to, what instead would the systemic rejection rate be for k randomly selected applicants were each applied to one of the positions? We find that the observed systemic rejection rate is always greater than the baseline systemic rejection rate, and that the discrepancy widens in relative terms as we consider applicants to more positions from 4 to 6 to 8 positions in Figure 2.

If each position selected applicants at the same rate and the recommendations for each position were made independently, then the systemic rejection rate would fall exponentially in the number of applications. In the final panel of Figure 2, we plot the observed and baseline systemic rejection rates as a function of the number of applications. We find that the observed systemic rejection rate follows an exponential, aligning with the basic theoretical intuition. From 2 to 8 applications, the disparity in systemic rejections between the the observed and baseline curves is roughly 5% for all values. When we fit both curves with exponential functions (observed = with a fit of ; baseline = with a fit of), we find that the coefficients of the exponentials are considerably different. In particular, while the baseline systemic rejection rate according to its exponential fit falls below 0.1% for applications, the observed systemic rejection rate only falls below this value at applications.

Our findings establish concrete evidence that deployed hiring algorithms in the wild yield significant homogeneity in outcomes. These findings are insufficient to establish that these applicants are rejected by the employers when they are not recommended (which we believe to be true), and do not account for the applicants' outcomes when applying to positions not mediated by pymetrics or the outcomes from alternative non-algorithmic decision-making. Therefore, building upon our findings, we encourage future work to explore these additional factors, addressing the practical hurdles in section 4 that currently make such analysis difficult, if not impossible.

Simulated outcomes. While most applicants apply to multiple positions, they never apply to *all* positions in the labor market. Candidate decisions about where to apply are often deliberate: applicants do not apply to jobs they don't desire, to jobs in distant geographies, or jobs in entirely

different market sectors. While the existing literature on employment posits many reasons for why applicants apply where they do [Kuhn and Shen, 2023, Marinescu and Rathelot, 2018], with some empirical work characterizing applicant behavior [Belot et al., 2018], no complete understanding of applicant behavior at economy scale exists currently.

Candidate outcomes are constrained by where they apply. Given our focus on the overall outcomes of candidates, especially whether they receive homogeneous outcomes and are systemically rejected, we explore how these outcomes depend on the choices candidates make. What would happen if, counterfactually, if applied to more positions? Would applicants avoid systemic rejection, potentially indicating that while employers may deem them a poor fit for the positions they applied to, there exists others where they would receive better outcomes?

Studying these outcomes is challenging for conventional methodologies in labor economics, especially those that rely on observational data, since the outcomes of interest are partially counterfactual. However, we capitalize on the unique possibilities for research made viable due to algorithmic hiring. By leveraging the centralized pymetrics data, the deterministic replicability of pymetrics hiring algorithms, and the efficiency of algorithmic decision-making, we simulate the outcomes for applicants as if they applied to every position mediated by pymetrics algorithms. Concretely, we sample 1000 applicants at random and have pymetrics evaluate them against each of their 555 models.[2]

Finding 3: Simulated outcomes reveal that all applicants would be recommended for at least one position, but that outcomes are frequent.

In Figure 3 (left), we depict how the real outcomes of applicants, associated with the sparse set of positions they applied to, relate to those we study in our simulation. Empirically, we find that none of the sampled 1000 applicants is rejected by every pymetrics models. Instead, of the 555 models that assess applicants, the applicant to be recommended the least receives recommendations with . In other words, while % of the sampled applicants were rejected from every position they applied to, these applicants would have received recommendations from the pymetrics algorithms if they had applied to every possible position.

Applying to every possible position is not only practically infeasible, but also does not account for the many factors that meaningfully shape applicants' decisions on where they apply. Given applicant outcomes for every position, we constrain where applicants apply to better reflect application patterns that would realize in the world. For each applicant, given the set of positions they applied to in reality, we expand this set to include additional positions that share applicants with the positions they applied to (see left of Figure 3). For the resulting set of positions, which for each applicant is necessarily a superset of where they actually applied, we plot the resulting distribution

4 Discussion

Hiring algorithms change lives and reconstitute labor markets, but we lack empirical research into their impact. Prior work on algorithmic hiring generally centers the perspective of firms. According to this perspective, the key factors to evaluate include the algorithm's validity and accuracy in reflecting the client firm's preferences, its scalability and efficiency in reducing the costs of hiring to the client firm, and its compliance with existing employment non-discrimination laws. Centering firms reflects standard practice in labor economics and the study of employment.

Yet given the impact that algorithmic hiring has on society, the firm-centric perspective is incomplete. In particular, the firm-centric perspective neglects the structural shift in the whole labor market imposed by the rise of algorithms. Algorithms influence the hiring processes at individual firms, but they also link the outcomes of these processes across firms due to pervasive algorithmic monoculture. And since job applicants typically apply to many jobs but (typically) only require one job, their outcomes are strongly influenced by the pattern of outcomes they receive across many employers. This means that to more comprehensively characterize the societal impact of algorithmic hiring, we must center the perspective of the job applicants. With this in mind, our work offers new evidence that we use to offer policy recommendations.

[2]Since these models were developed over several years by pymetrics, during which they changed their infrastructure, only applicants were able to evaluated against every model.

Recommendation 1: Evaluate employment discrimination on a per-position level.

Recommendation 2: Track homogeneous outcomes in hiring. We demonstrate that algorithmic hiring in the case of pymetrics results in homogeneous outcomes for individual applicants more often than in our baseline. However, we emphasize that while this outcome may be visible to pymetrics and to individual applicants, it is not visible to firms nor is it tracked by regulators such as the U.S. Equal Employment Opportunity Commission. Given that employment is an essential good for individual welfare [Walzer, 1983, Anderson, 2017], there is a strong social imperative to understand homogeneous outcomes given they may directly correspond with systemic exclusion from the labor market or unemployment. We recommend that governments invest in mechanisms for monitoring the prevalence of homogeneous outcomes and the extent to which they are driven by algorithmic hiring. These mechanisms could relate to existing mechanisms for monitoring the unemployment rate and length of job searches. If systemic rejection from jobs mediated by algorithmic hiring leads a candidate to be unemployed, or unemployed for longer than they would otherwise be, the length of unemployment can be a cascading harm, decreasing future likelihood of receiving an interview [Kroft et al., 2013b, Farber et al., 2019].

Recommendation 3: Track model-sharing and data-sharing across employers. Algorithmic monoculture [Kleinberg and Raghavan, 2021, Jagadeesan et al., 2023] is on the rise across many domains. Each model creator has an incentive to rely on the best data, models, and model architecture available, leading to concentration in approaches and outcomes, as when tens of thousands of developers around the world rely on OpenAI's language models [OpenAI and Pilipiszyn, 2021].

We believe algorithmic monoculture has unique implications in hiring. In other aspects of employment, competitive practices are seen as essential, prompting regulation to prohibit anti-competitive practices such as collusion to set wages across firms. By contrast, we see multiple direct competitors in very narrowly defined industries using pymetrics's services. We encourage policymakers to question whether it is acceptable for firms, especially those that compete with each other, to adopt the same algorithms or models in their hiring practices.

In particular, we highlight two concerns: *data sharing* across firms and *mechanical homogenization*.

Recommendation 4: Require transparency. Companies must inform candidates about use of algorithmic hiring and reliance on shared models. At present, candidates are often unaware of the presence of algorithmic hiring.[3] Given the consequences of algorithmic hiring, we recommend candidates be notified of the use of hiring algorithms. This recommendation is concordant with proposed (but not current) U.S. legislation. The OSTP's Blueprint for an AI Bill of Rights includes rights regarding notice and explanation, including, "You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you" [OSTP, 2022]. New York's Local Law 144 includes a section on notification (§5-304 Notice to Candidates and Employees), as do California's proposed AB 331 and New York State's proposed NY A00567, but it is unclear to which algorithmic hiring processes these laws will be applied. New Jersey's proposed A4909 would apply to third party AEDTs (automated employment decision tools) that filter candidate applications. However, beyond simply being made aware of the presence of algorithms, we believe it is critical that candidates are clear about how outcomes are linked across employers. Consistent with work on dark patterns, deceptive practices, and other forms of mismanaged consumer expectations in the presence of algorithms, it is inappropriate for candidates to believe certain outcomes are possible when in fact they are not under algorithmic conditions. In this context, candidates should not believe they can receive different outcomes across employers if this is impossible. Namely, if two employers (especially competitors) made decisions on the basis of exact same machine learning model or algorithmic process, candidates should minimally be aware this that is possible. Notification creates the possibility of algorithmic recourse should this coordination be judged to be inappropriate.

Recommendation 5: Mandating access for researchers into algorithmic hiring processes. Our work demonstrates the value of empirical work conducted by external researchers in advancing public understanding of algorithmic hiring. To date, empirical work on hiring has been made possible

[3]When candidates are aware that algorithmic hiring is occurring, studies have suggested that candidates tend to believe it will be more accurate in assessing their performance, even without evidence attesting this is true of the particular algorithm by which they are being evaluated Fumagalli et al. [2022].

through a variety of study designs. Perhaps most famously, *correspondence studies* or audit studies have established biases in several parts of the economy [Bertrand and Mullainathan, 2004]. While such designs have received some critique [Cahuc et al., 2019], we emphasize that correspondence study approaches are infeasible for understanding the role of algorithmic hiring in the labor market for a different reason: they evaluate the outcomes of the hiring process, which leaves unclear whether the outcomes are algorithmically-mediated and if so, at which stages of hiring algorithms appear. Research on algorithmic hiring has studied publicly accessible job posting platforms such as Indeed and LinkedIn [Zhang and Kuhn, 2022], but many hiring algorithms and associated job outcomes are neither accessible to the public nor to researchers. pymetrics is unusually transparent, having allowed researchers access to data for this study as well as prior works [Wilson et al., 2021, Kassir et al., 2023].We encourage heightened transparency into these pipelines for researchers, especially to the extent that hiring algorithms only play a partial role in determining the ultimate job outcomes for individuals, and especially given that hiring algorithms mediate high-stakes decisions at population scale.

5 Methods

Data acquisition. To acquire data, the researchers negotiated access to pymetrics data that spans applicant gameplay features, applicant demographics, firm metadata, and algorithmic recommendations that pymetrics makes for each application (i.e. the percentile $p \in [0, 1]$ and the thresholded binary recommended $y \in \{0, 1\}$). Data access is subject to a non-disclosure agreement that gives researchers' insight into the exact methods that generate the data, which include proprietary information. Moreover, the researchers signed a data-use agreement that: (i) facilitates secure transfer of pymetrics data from pymetrics to the researchers, (ii) restricts redistribution of data by the researchers, and (iii) requires the research manuscript be provided to pymetrics one month in advance of publication. All aspects of the research: conceptualization, design, execution, analysis, and presentation are conducted by the researchers independent of pymetrics, with no conflicts of interest. The exclusive role of pymetrics is to provide the required data for this research, provide clarification on internal practices (e.g. best practices for data processing), and to limit disclosure of any proprietary information. Beyond the agreements between pymetrics and the researchers, the research was subject to IRB review (IRB protocol 68164) as well as a data risk assessment by the research institution.

Data processing. Given the acquired data from pymetrics, we process the data in four steps. First, we remove test models based on information from pymetrics as these do not correspond to real applications. Second, we remove applications with missing predicted job fit percentile scores, because they lack the outcome of interest. Third, we de-duplicate applications that have the same user ID, position ID, organization ID and model ID, limiting each user to one application to a specific position. We keep the application with the most demographic information that is the oldest. Finally, we resolve an inconsistency to facilitate data processing: in five cases, two employer organizations were assigned the same organization ID which we correct so each organization has a unique organization ID.

Bias metric. To measure bias, we use well-known guidance from the U.S. Equal Employment Opportunity Commission. For each position, we consider applicants with self-reported demographic information and compute the selection rate for each demographic group. To measure bias for a given demographic group g in the context of a fixed position, denote the group's selection rate s_g. Denote the most selected group as g' and its selection rate as $s_{g'}$. The *impact ratio* r_g for g is defined as $\frac{s_g}{s_{g'}}$. The *test statistic* z_g for g is the output of a two-sample pooled-proportion z-test with inputs s_g and $s_{g'}$ along with the sizes of g and g'. g experiences *adverse impact* if (i) $r_g < 0.8$ **and** (ii) $z_g > 1.96$.

Homogeneity baseline. To contextualize the observed rate of systemic rejection, we compare the observed rate to a baseline. In contrast to the observed rate, which considers how often an applicant is not recommended for all k positions they applied to, the baseline considers the rate at which k randomly chosen applicants would not be recommended for all k positions if each applicant assigned to one position. Formally, let N applicants each apply to k positions with $h_1, \ldots, h_k$ denoting the associated decision functions.[4] For applicant i, let x_i denote their gameplay features and $\hat{y}_i^j = h_j(x_i)$

[4]The above notation (for simplicity) assumes all applicants apply to the same k positions: we generalize the notation in the supplement to address applicants applying to differing sets of k positions.

be their binary outcome. Let $\hat{y}_i^j = 1$ indicate being recommended, whereas 0 indicates being not recommended. Define the outcome matrix $O \in \{0,1\}^{N \times k}$ such that $O[i,j] = \hat{y}_i^j$. For each position j, define the selection rate $s_j = \frac{\sum_{i=1}^{N} O[i,j]}{N}$.

For $t \in \{0, \ldots, k\}$, the observed rate at which individuals receive t recommendations, and the baseline rate for t recommendations, are defined as follows:

$$P_{\text{observed}}(t \text{ rec.}) = \frac{\sum_{i=1}^{N} \mathbb{I}\left[t = \sum_{j=1}^{k} O[i,j]\right]}{N} \tag{1}$$

$$P_{\text{baseline}}(t \text{ rec.}) = \text{Poisson-Binomial}(s_1, \ldots, s_k)[t] \tag{2}$$

The observed and baseline systemic rejection rates are $P_{\text{observed}}(0)$ and $P_{\text{baseline}}(0)$, respectively.

Simulation data. To generate the simulated outcomes, we provided pymetrics with three samples of 1000 applicants. These samples were constructed based on the following criteria:

1. A purely random sample drawn from the entire applicant pool.
2. A stratified random sample consisting of only applicants who provided a location and were from the United States.
3. A stratified random sample of applicants with multiple applications, specifically: 100 applicants randomly selected from each subset of applicants with i applications, where i ranges from 2 to 11, inclusive.

pymetrics evaluated the selected applicants using all applicable pymetrics models, generating a percentile score for each applicant-model pair. The resulting dataset is structured at the applicant-by-model level. For a small set of applicants, we lack scores for certain models, often due to misalignment between the input data features and the features the version of the model expects. On average, each applicant in each of the three samples was evaluated by TOFIND models. We use the the purely random sample for most simulation data analyses and the stratified U.S.-based sample for the additional bias-related analyses.

Simulation analysis methods. We generate the connected set, described in the results section. For our first simulation, we selected users whose connected set includes at least 25 models and randomly sampled exactly 25 model interactions per user to ensure consistent comparison. We then performed Monte Carlo simulations (n=100) to compare the observed distribution of rejections against the baseline distribution expected if model rejections were independent. This approach allowed us to quantify homogenization amongst a random sample of users who interact with a large number of models.

For our second set of simulations, we extended our previous analysis by examining the relationship between the number of models in a user's connected set and the probability of experiencing rejection. While our first simulation focused on exactly 25 models per user, we now repeated the exercise for different numbers of models, simulating user interactions with 2 to 100 models. This approach helped us determine at what point systemic rejection disappears. The threshold at which users in the random sample no longer experience systemic rejection is identified as the point where the confidence interval for the mean number of rejections includes zero.

References

J. Fuller, M. Raman, E. Sage-Gavin, and K. Hines. Hidden workers: Untapped talent. Technical report, Harvard Business School Project on Managing the Future of Work and Accenture, 9 2021.

David Autor and David Scarborough. Does job testing harm minority workers? evidence from retail establishments. *The Quarterly Journal of Economics*, February 2008.

Mitchell Hoffman, Lisa B Kahn, and Danielle Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. Algorithmic pluralism: A structural approach to equal opportunity. In *The 2024 ACM Conference on Fairness, Accountability,*

and Transparency, FAccT '24. ACM, June 2024. doi: 10.1145/3630106.3658899. URL http://dx.doi.org/10.1145/3630106.3658899.

Allie Nawrat. Inside hirevue's acquisition of modern hire, May 2023. URL https://www.unleash.ai/hr-technology/inside-hirevues-acquisition-of-modern-hire/. Accessed on March 25, 2024.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2018340118. URL https://www.pnas.org/content/118/22/e2018340118.

Rishi Bommasani, Kathleen Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture homogenize outcomes? In *Advances in Neural Information Processing Systems*, 2022.

Kathleen Creel and Deborah Hellman. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, 52(1): 26–43, 2022. doi: 10.1017/can.2022.3.

Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT 2021)*, Virtual Event, Canada, March 2021.

Sara Kassir, Lewis Baker, Jackson Dolphin, and Frida Polli. Ai for hiring in context: a perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI Ethics*, (3):845–868, 2023. doi: https://doi.org/10.1007/s43681-022-00208-x.

Ifeoma Ajunwa. An auditing imperative for automated hiring. *Harvard Journal of Law and Technology*, (1), 2021.

Louis S. Jacobson, Robert J. LaLonde, and Daniel G. Sullivan. Earnings losses of displaced workers. *The American Economic Review*, 83(4):685–709, 1993. ISSN 00028282. URL http://www.jstor.org/stable/2117574.

Amartya Sen. Inequality, unemployment and contemporary europe. *Int'l Lab. Rev.*, 136:155, 1997.

Daniel Sullivan and Till Von Wachter. Job displacement and mortality: An analysis using administrative data. *The Quarterly Journal of Economics*, 124(3):1265–1306, 2009.

Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4): 991–1013, August 2004. doi: 10.1257/0002828042002561. URL https://doi.org/10.1257/0002828042002561.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230. URL https://www.science.org/doi/abs/10.1126/science.aal4230.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, 2024.

Michael R. Dalton and Jeffrey A. Groen. How do jobseekers search for jobs? new data on applications, interviews, and job offers. *Beyond the Numbers: Employment & Unemployment*, 9(14), November 2020. URL https://www.bls.gov/opub/btn/volume-9/how-do-jobseekers-search-for-jobs.htm.

Kory Kroft, Fabian Lange, and Matthew J. Notowidigdo. Duration dependence and labor market conditions: Evidence from a field experiment*. *The Quarterly Journal of Economics*, 128(3): 1123–1167, 06 2013a. ISSN 0033-5533. doi: 10.1093/qje/qjt015. URL https://doi.org/10.1093/qje/qjt015.

Donna S. Rothstein. An analysis of long-term unemployment. *Monthly Labor Review*, July 2016. URL https://www.bls.gov/opub/mlr/2016/article/an-analysis-of-long-term-unemployment.htm.

Chang-Tai Hsieh, Erik Hurst, Charles I Jones, and Peter J Klenow. The allocation of talent and us economic growth. *Econometrica*, 87(5):1439–1474, 2019.

Peter Kuhn and Kailing Shen. What happens when employers can no longer discriminate in job ads? *American Economic Review*, 113(4):1013–48, April 2023. doi: 10.1257/aer.20211127. URL https://www.aeaweb.org/articles?id=10.1257/aer.20211127.

Ioana Marinescu and Roland Rathelot. Mismatch unemployment and the geography of job search. *American Economic Journal: Macroeconomics*, 10(3):42–70, July 2018. doi: 10.1257/mac.20160312. URL https://www.aeaweb.org/articles?id=10.1257/mac.20160312.

Michèle Belot, Philipp Kircher, and Paul Muller. Providing advice to jobseekers at low cost: An experimental study on online advice. *The Review of Economic Studies*, 86(4):1411–1447, 10 2018. ISSN 0034-6527. doi: 10.1093/restud/rdy059. URL https://doi.org/10.1093/restud/rdy059.

Michael Walzer. *Spheres of Justice: A Defense of Pluralism and Equality*. Basic Books, 1983.

Elizabeth Anderson. *Private Government: How Employers Rule Our Lives (and Why We Don't Talk about It)*. Princeton University Press, 2017.

Kory Kroft, Fabian Lange, and Matthew J Notowidigdo. Duration dependence and labor market conditions: Evidence from a field experiment. *The Quarterly Journal of Economics*, 128(3):1123–1167, 2013b.

Henry S. Farber, Chris M. Herbst, Dan Silverman, and Till von Wachter. Whom do employers want? the role of recent employment and unemployment status and age. *Journal of Labor Economics*, 37(2):323–349, 2019. doi: 10.1086/700184. URL https://doi.org/10.1086/700184.

Meena Jagadeesan, Michael I. Jordan, Jacob Steinhardt, and Nika Haghtalab. Improved bayes risk can yield reduced social welfare under competition, 2023.

OpenAI and Ashley Pilipiszyn. Gpt-3 powers the next generation of apps. https://openai.com/blog/gpt-3-apps, 3 2021. Accessed: 2023-06-24.

Elena Fumagalli, Sarah Rezaei, and Anna Salomons. Ok computer: Worker perceptions of algorithmic recruitment. *Research Policy*, 51(2):104420, 2022. ISSN 0048-7333. doi: https://doi.org/10.1016/j.respol.2021.104420. URL https://www.sciencedirect.com/science/article/pii/S0048733321002146.

White House OSTP. Notice and explanation. https://www.whitehouse.gov/ostp/ai-bill-of-rights/notice-and-explanation/, 10 2022. Accessed: 2023-06-24.

P. Cahuc, S. Carcillo, A. Minea, and M.-A. Valfore. When correspondence studies fail to detect hiring discrimination. Discussion Paper 14028, CEPR, 2019.

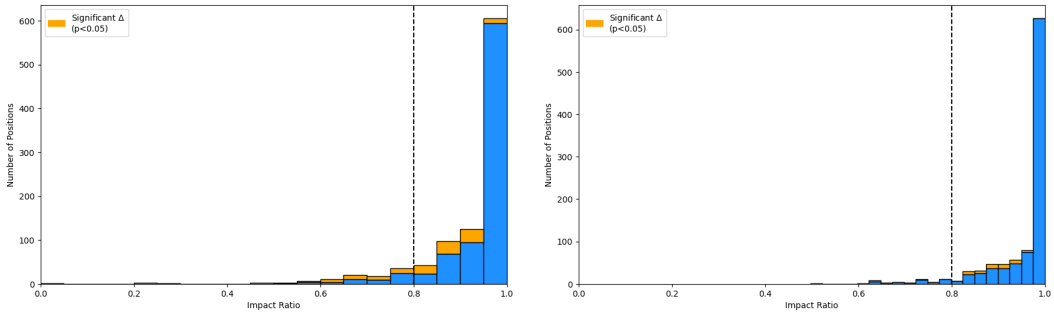
Shuo Zhang and Peter Kuhn. Understanding algorithmic bias in job recommender systems: An audit study approach. 2022.

A Additional information on data

B Additional information on bias analyses

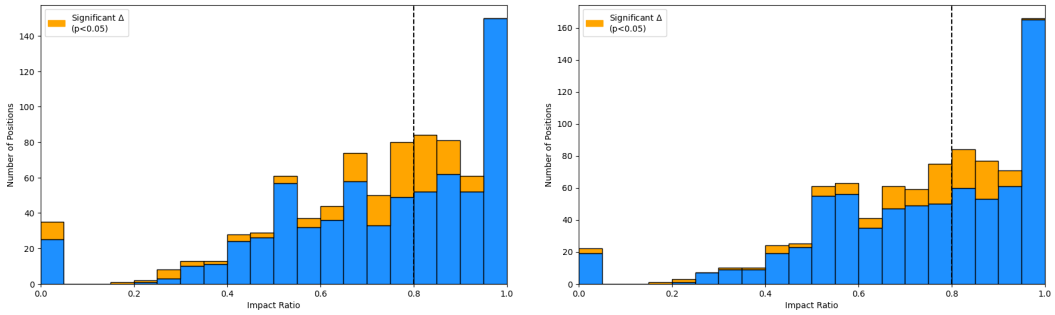
C Additional information on homogeneity analyses

D Additional information on simulation analyses



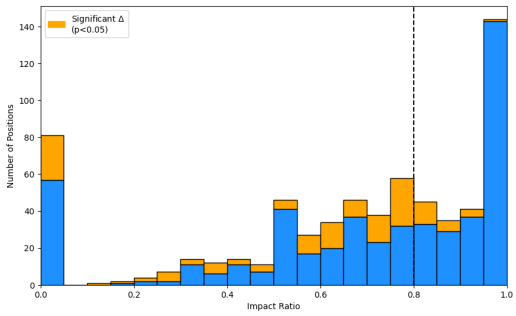
(a) Female (b) Male

Figure 4: **Gender bias in selection rates.** The histogram for each gender group corresponding to the impact ratio for the group for each of the disaggregated positions. Emphasis is placed on impact ratios below 0.8 and where the group's selection rate is significantly different from that of the most-selected group for $p < 0.05$.

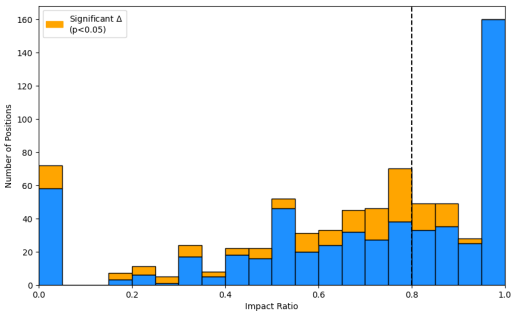


(a) Female Asian

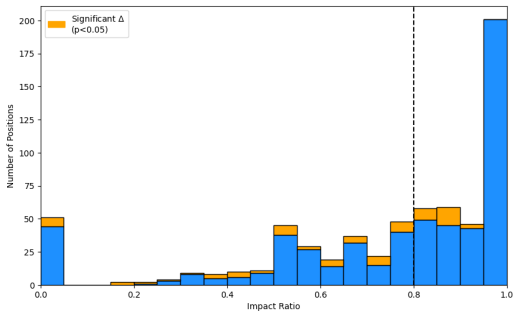
(b) Male Asian



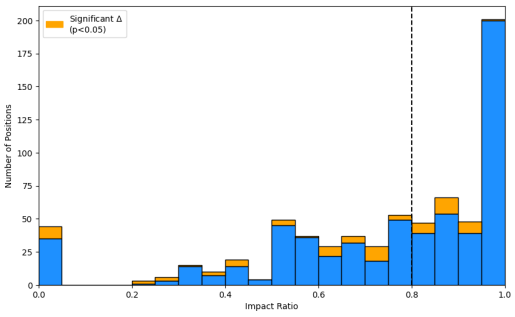
(c) Female Black



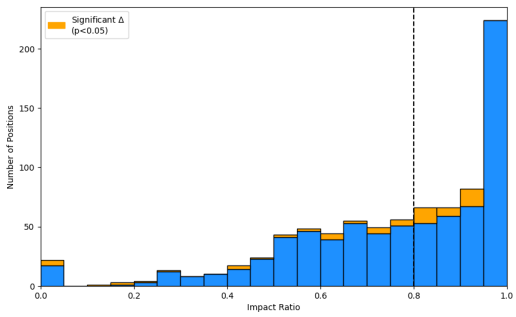
(d) Male Black



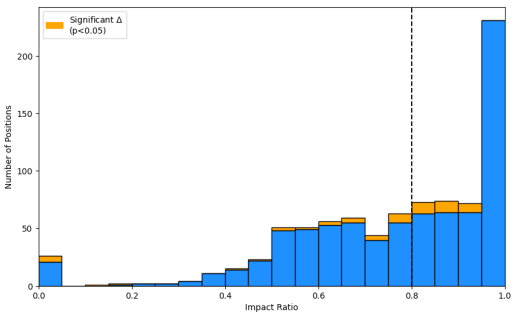
(e) Female Hispanic/Latino



(f) Male Hispanic/Latino



(g) Female White



(h) Male White

Figure 5: **Intersectional bias in selection rates.** The histogram for each racial group corresponding to the impact ratio for the group for each of the disaggregated positions. Emphasis is placed on impact ratios below 0.8 and where the group's selection rate is significantly different from that of the most-selected group for $p < 0.05$.