Algorithmic Monocultures in Hiring

RISHI BOMMASANI*, Stanford University, USA
SARAH H. BANA*, Chapman University, USA
KATHLEEN A. CREEL*, Northeastern University, USA
DAN JURAFSKY, Stanford University, USA
PERCY LIANG, Stanford University, USA

Many employers screen job applicants with algorithms built by the same few algorithm vendors. We hypothesize that algorithmic monoculture leads to the same individuals and members of the same racial groups facing rejection. We acquire and analyze a novel dataset of 3 million applicants submitting 4 million applications where all the applications are screened by algorithms built by the same vendor. We find clear racial disparities in applicant outcomes. Of all applications submitted by Asian and Black applicants, 14.74% and 25.87% are submitted to positions that adversely impact Asian and Black applicants, respectively, according to U.S. employment discrimination standards. Individuals also receive homogeneous outcomes: 4% of all applicants who apply to 10 positions are recommended for rejection from all positions, a rate higher than expected by chance. To better understand this homogeneity, we leverage the deterministic replicability of hiring algorithms to generate the outcomes applicants would have received if they applied to all positions. We show that applicants would need to apply widely in order to ensure their applications are considered by a human.

CCS Concepts: • **Computing methodologies** → **Machine learning**; **Artificial intelligence**; • **Applied computing** → **Consumer products**.

Additional Key Words and Phrases: algorithmic monoculture, outcome homogenization, algorithmic fairness, algorithmic hiring, algorithmic screening

ACM Reference Format:
Rishi Bommasani, Sarah H. Bana, Kathleen A. Creel, Dan Jurafsky, and Percy Liang. 2026. Algorithmic Monocultures in Hiring. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 32 pages. https://doi.org/10.1145/3805689.3812400

1 Introduction

Over 90% of U.S. employers rely on hiring algorithms to screen or rank job applicants [20]. Hiring algorithms shape which applicants are considered for an interview and which applications are never seen by a human [3, 23, 42]. They are a bottleneck to opportunity [29] for billions of workers. Many employers procure hiring algorithms from the same few third-party vendors. As of May 2023, over 60% of the Fortune 100 and eight of the ten largest US federal agencies use HireVue's algorithms [44]. By mediating screening for multiple employers, hiring algorithms establish an *algorithmic monoculture*, defined as the state in which many decision-makers rely on the same or similar algorithms [10, 37]. We hypothesize that algorithmic monoculture leads to homogeneous outcomes:

*All three authors contributed equally to this research.

Authors' Contact Information: Rishi Bommasani, Stanford University, Stanford, CA, USA; Sarah H. Bana, Chapman University, Orange, CA, USA; Kathleen A. Creel, Northeastern University, Boston, USA; Dan Jurafsky, Stanford University, Stanford, CA, USA; Percy Liang, Stanford University, Stanford, CA, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812400

encountering these algorithms repeatedly will result in the *same groups* experiencing adverse impact and the *same individuals* being rejected [1, 10, 15].

To understand algorithmic hiring in practice, we conduct the first study observing deployed algorithmic hiring decisions across multiple employers from a single vendor. We acquire a novel dataset from the talent platform pymetrics, which records 4,197,168 job applications submitted by 3,372,132 applicants to 1,746 positions.[1] Figure 1 shows the pymetrics-mediated hiring pipeline. Each application is assessed by a pymetrics machine learning model that assigns a score that is binarized into outcomes of "recommend" or "do not recommend." pymetrics's recommendations inform their clients' decisions about which applicants to interview, reject, and hire. When pymetrics's algorithms "do not recommend" an applicant, they are likely to be rejected without consideration by a human [20]. Our dataset spans recommendations that influence hiring at 156 employers with a cumulative annual revenue of $225 billion dollars across 11 industries including finance, manufacturing, and warehousing.

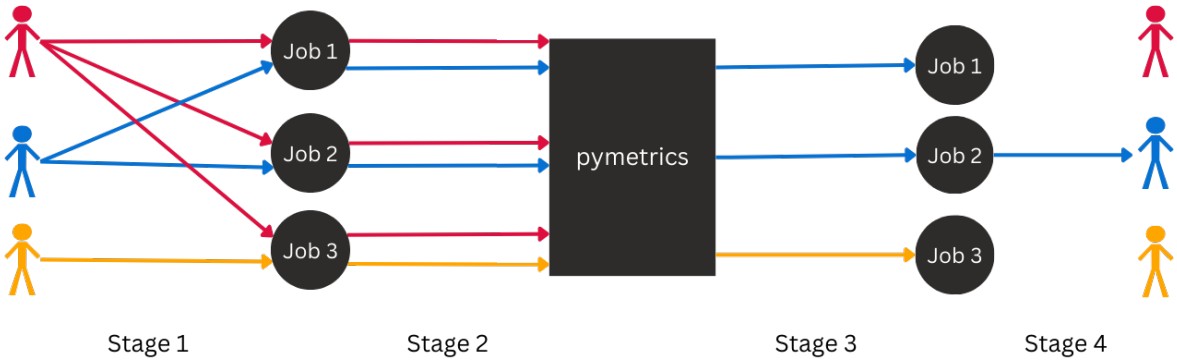


Fig. 1. **The pymetrics process.** Stage 1: Applicants apply to positions. Stage 2: Applicants are directed to the pymetrics platform to play assessment games. Stage 3: pymetrics algorithms use applicant gameplay features to recommend 58.2% of applicants per position on average. Stage 4: Employers decide which applicants to interview or hire, typically rejecting applicants that were not recommended by pymetrics.

We find evidence of adverse impact based on race. In a previous analysis of selection rates for many demographic groups [34], researchers from pymetrics found no differences that would prompt scrutiny according to U.S. employment discrimination law. Investigation into substantial differences can occur when the "impact ratio," namely the ratio between the selection rates of the most selected group and the group of interest, is less than 0.8 and statistically significant, a standard colloquially referred to as the "4/5ths rule." However, this study reported only aggregate results across all applications, irrespective of the position or employer they are associated with. Since U.S. guidelines operationalize discrimination on a per-job basis [13, 41 CFR 60-3.15.2(a)], we instead study each of the 1,746 positions separately. Disaggregating on a per-position basis reveals that 10.62% of positions demonstrate adverse impact against Black applicants. 30.70% of Black applicants apply to at least one position that adversely impacts Black applicants and 25.87% of applications submitted by Black applicants are to models with adverse impact against Black applicants.

Our findings support the view that hiring algorithms can demonstrate adverse impact. Prior work finds discriminatory patterns in decisions based on applicant resumes [7, 21, 47, 48, 56] that include demographic proxies such as racialized names like Jamal [7] or gender-stereotyped activities like softball [5]. In contrast, pymetrics

[1]The data is provided for independent research: pymetrics is unable to edit, restrict, or veto the research as stated in the data use agreement (Appendix A). This is consistent with recommendations made by [62].

screens applicants based on their performance on online games. Game-based and video-based assessments are among the most common and fastest-growing forms of algorithmic screening [49]. We find adverse impact even though these games do not overtly embed demographic information and even though pymetrics aims to pro-actively debias each model they produce [34, pg. 859]. [2] Our findings build on prior work showing that AI can have discriminatory effects even in the absence of explicit demographic information [4, 12, 24], as in "proxy discrimination" when predictive models discriminate by relying on variables that are proxies for demographic information in making their predictions [26, 32, 33].

We also provide the first evidence of systemic rejections at scale due to deployed hiring algorithms. Prior work [1, 10, 15, 37] theorizes that monoculture could lead to an applicant being rejected everywhere they apply, as is the case for the red applicant in Figure 1. 42 pymetrics models screen applicants at multiple companies, which means a rejection at one company mechanically entails rejection from another company using the same model. Even when applicants are evaluated by different pymetrics models, empirically some are systemically rejected. Of applicants that apply to ten positions, 4% are rejected from all positions. As applicants apply to more positions, we find that the systemic rejection rate falls exponentially ($R^2 = 0.984$): while this exponential decay would be predicted even if decisions were made independently, the rate itself decays more slowly than would be expected by chance. To the best of our knowledge, this pattern of systemic rejection is distinctive to algorithmic hiring. Analyzing data from the largest prior study of hiring decisions [38], which sent 83,000 applications to 108 Fortune 500 firms, we find that the systemic rejection rates observed in their data are very accurately predicted by employers making statistically independent decisions. Our data systematically diverges from this baseline of independence, showing that algorithmic monoculture produces qualitatively different labor market dynamics.

Data access limits empirical research on algorithmic hiring. Single-employer studies [14, 58] cannot observe whether rejection at Firm A predicts rejection at Firm B because they lack cross-employer applicant tracking, correspondence studies [7, 38] cannot observe the underlying decisionmaking process to study the role of algorithms, and static audits of hiring algorithms [60] cannot observe deployment-time outcomes. Our dataset is the first to observe real algorithmic outcomes for the same applicant across multiple employers. The algorithmic setting also allows us to study new research questions: we introduce new methods that capitalize on the deterministic replicability of algorithmic recommendations and the efficiency of algorithmic decision-making to answer a research question that would be intractable for traditional social science methods. We generate the counterfactual outcomes applicants would receive if they applied to every position and were, thereby, assessed by every pymetrics model. This simulation allows us to determine whether candidates are systemically rejected only because they apply to jobs for which they would be a poor fit, such that if they applied to more or different jobs they would be accepted. Our simulation shows that every applicant would be recommended by at least one pymetrics model, even though in reality many applicants were not recommended for any position to which they applied. Under more realistic applicant behavior, where applicants apply more broadly but not everywhere, we find that some applicants are still systemically rejected. To guarantee a systemic rejection rate below 0.1%, applicants would need to submit 25 applications compared to 10 under the baseline of independence.

Independent research is necessary to illuminate otherwise-opaque hiring algorithms. By consolidating part of hiring decision process across distinct employers, hiring algorithms impact collective adverse impact rates and patterns of systemic rejection, which are particularly salient given the established harms of discrimination and extended unemployment [28, 52, 55]. As algorithmic hiring policy advances (e.g. New York City Local Law 144 of 2021, the EU AI Act of 2024), we recommend that policymakers work to increase transparency into algorithmic hiring and create new pathways for independent research.

[2] External researchers in a pymetrics-funded study supported this claim, finding that the seven models they inspected were constrained to avoid adverse impact during training [60, pg. 4].

2 Data

We analyze data from the hiring algorithm vendor pymetrics from December 2018 through December 2022.[3] We maintain full independence in research design, analysis, and interpretation with data access granted until December 31, 2025 subject to non-disclosure agreements regarding proprietary company information (see Appendix A). Our data involves job-seeking applicants, job-providing employers, and pymetrics. Clients of pymetrics direct their job applicants to the pymetrics platform, where applicants play 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 use machine learning to classify applications. pymetrics's 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 2,404 applications on average. Almost 60% of all applications are to employers in four major industries (professional services, financial services, manufacturing, and technology) for positions that are primarily white-collar with the notable exception of Hand Laborers and Material Movers (12.67% of applications). The majority of the 156 employers are located in North America and the majority of the employers have annual revenues of at least $5 billion.

pymetrics. pymetrics builds 16 online games to measure applicants' cognitive traits, including propensity to take risks, processing speed, trust, altruism, and planning ability. For each client, pymetrics trains a binary classifier: positive training examples correspond to the gameplay features of at least 50 current employees in that role and negative training examples correspond to the gameplay features of random profiles in the pymetrics database [46].[4] Choosing which employees will serve as the positive examples is the primary way that the employer influences the classifier.

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

Recommendations. When an applicant applies for 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 below 0.5 yield predictions $y = 0$ ("do not recommend") and scores above 0.5 yield predictions $y = 1$ ("recommend").[5] On average, 41.8% of applications are "not recommended", meaning that a hiring manager is unlikely to further consider them. We consider this to be equivalent to rejection in the high-volume hiring context

[3] In August 2022, pymetrics was acquired.

[4] Prior work has critiqued the use of game-play features for job candidate screening, including critiquing the stability of psychometric personality tests [50], the assumption that game-play scores can be a predictor of job success [53], and the many construct reliability and validity concepts that could be used to assess the efficacy of measurement [27]. Our work does not measure the validity of the gameplay features for prediction.

[5] This method for converting the scores into binary outcomes is the practice pymetrics uses when they publish results [34] and the practice they recommended to us. However, in practice, pymetrics recommendations are used in different ways by different clients. Some clients prefer the "recommend" category to be divided into two categories, denoted by "recommend" and "highly recommend" (sometimes color-coded as a ternary red-yellow-green system in an applicant tracking software user interface). Other options include quintiles where recommendations are divided into five tiers. Throughout this work, we use the binary "recommend" and "do not recommend" categories, consistent with the analysis practices and guidance from pymetrics. Since our focus is primarily on rejections (i.e. "do not recommend"), the finer divisions that sometimes subdivide the "recommend" category from the client's perspective do not affect our analysis.

in which pymetrics operates. Since each position receives an average of 2,404 applications, it is not feasible for humans to manually review all applications.[6]

Sample. Our data tracks 4,197,168 applications. It includes applicant gameplay features and for each application, the application date, the position name and employer, metadata about the position and employer, and the numerical score and final recommendation each applicant received for each completed application. 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). Employers who use pymetrics may have already collected demographic data when an applicant initially applies. Employers decide whether or not to allow pymetrics to collect demographic data. Because each applicant has a unique ID, we can track the same applicant as they apply to distinct positions across the study time period. The majority of applicants submit just one application to a pymetrics-mediated position, though over five hundred thousand applicants submit multiple applications, which totals to 1.2 million applications.

Table 1. **Self-Reported Applicant Descriptive Statistics**.

Gender		*Race*		*Country*	
Male	28.31	Asian	16.82	United States	16.05
Female	19.47	White	14.23	India	6.56
Prefer Not to Say	0.17	Black	3.57	United Kingdom	3.26
Other	0.07	Hispanic/Latino	3.02	Australia	2.03
Missing	51.97	Other	2.55	Other	17.70
		Missing	59.83	Missing	54.39

The data is tabular with rows corresponding to applications and columns corresponding to application metadata (e.g. submission time), applicant metadata (e.g. race), position metadata (e.g. position name), employer metadata (e.g. employer name), model metadata (e.g. model ID), and the percentile scores $p \in [0, 1]$ that are the outputs of pymetrics models. As the first independent large-scale empirical study on algorithmic hiring, we elect to minimally process the data to maximize fidelity instead of smoothing away outliers or other anomalies for cleaner analyses. Data processing involves (i) removing test models that did not evaluate real applicants, (ii) removing unscored applications, (iii) deduplicating essentially identical applications and (iv) fixing coding errors where multiple employers were assigned the same organization ID. These decisions had minimal effects on the data and were discussed with data scientists at pymetrics.

3 Results

We study the data using three lenses: (i) adverse impact, a concern for hiring processes of all kinds; (ii) systemic rejection, a particular consideration for algorithmic hiring; and (iii) unique possibilities enabled by algorithmic hiring.

3.1 Adverse impact in algorithmic hiring

We measure adverse impact according to the guidance of the U.S. Equal Employment Opportunity Commission (EEOC): the EEOC is the federal agency tasked with enforcing employment discrimination in the United States [17]. pymetrics acknowledges the EEOC standard as the relevant standard [46]. Adverse impact occurs when there is (i) practically *and* (ii) statistically significant disparities in the *selection rate* s_g for the group of interest

[6]When dealing with fewer applications, prior work finds that applicants ranked in the bottom half can be short-listed between 18–31% of the time [18]. For participant interviews comparing high and low-volume hiring processes that rely on algorithmic decision-making systems, see [54].

Table 2. **Disparate Impact Ratio Results by SOC Major Occupation Group and Race**

SOC Major Occupation Group	Race	Impact Ratio	Positions	#Adverse Impact	%Adverse Impact	%Adverse Impact-BH
Management (11)	Asian	0.897	36	0	0.0	0.0
	Black	0.873	34	9	26.5	14.7
	Hispanic	0.871	38	2	5.3	0.0
	White	0.956	39	0	0.0	0.0
Business and Financial Ops (13)	Asian	0.876	140	13	9.3	7.1
	Black	0.834	115	24	20.9	13.9
	Hispanic	0.912	125	6	4.8	0.8
	White	0.979	140	0	0.0	0.0
Computer and Mathematical (15)	Asian	0.847	61	8	13.1	11.5
	Black	0.809	48	15	31.2	16.7
	Hispanic	0.883	38	5	13.2	2.6
	White	0.990	53	2	3.8	1.9
Architecture and Engineering (17)	Asian	0.790	10	1	10.0	10.0
	Black	0.646	8	2	25.0	25.0
	Hispanic	0.923	9	1	11.1	0.0
	White	0.986	12	0	0.0	0.0
Legal (23)	Asian	0.816	12	1	8.3	8.3
	Black	0.821	11	3	27.3	18.2
	Hispanic	0.912	12	0	0.0	0.0
	White	0.859	12	0	0.0	0.0
Sales and Related (41)	Asian	0.874	68	3	4.4	1.5
	Black	0.937	64	5	7.8	3.1
	Hispanic	0.914	71	2	2.8	0.0
	White	0.951	82	1	1.2	0.0
Office and Administrative Support (43)	Asian	0.810	43	4	9.3	4.7
	Black	0.828	41	5	12.2	7.3
	Hispanic	0.923	42	2	4.8	0.0
	White	0.981	43	1	2.3	0.0
Other SOC Code	Asian	0.927	28	3	10.7	3.6
	Black	0.843	26	2	7.7	0.0
	Hispanic	0.921	23	0	0.0	0.0
	White	0.906	27	0	0.0	0.0
No SOC Code	Asian	0.875	485	42	8.7	4.9
	Black	0.838	425	71	16.7	10.4
	Hispanic	0.920	390	28	7.2	1.0
	White	0.957	469	13	2.8	0.6
All	Asian	0.870	883	75	8.5	**5.3**
	Black	0.839	772	136	17.6	**10.6**
	Hispanic	0.916	748	46	6.1	**0.8**
	White	0.962	877	17	1.9	**0.5**

For each SOC Major Occupation Group and racial group, the table reports the aggregate impact ratio across all positions, the number of positions, and the number/share of positions that demonstrate adverse impact. We also report the share of positions that demonstrate adverse impact subject to the Benjamini–Hochberg correction (BH) threshold for $\alpha = 0.05$. Positions are included only if at least 30 applicants self-report race.

g when compared against the selection rate $s_{g'}$ of the most selected group g'.[7] Practical significance requires the *impact ratio* $r_g = \frac{s_g}{s_{g'}}$ to be less than 0.8, which is why the EEOC guidance is colloquially referred to as the "four-fifths" rule. Statistical significance requires the test statistic z_g to be at least 1.96, where z_g is the output of a two-sample pooled-proportion z-test with inputs s_g and $s_{g'}$ along with the number of applicants for the two groups (n_g, $n_{g'}$).

$$\hat{p} = \frac{s_g n_g + s_{g'} n_{g'}}{n_g + n_{g'}}$$

$$z_g = \frac{s_g - s_{g'}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_g} + \frac{1}{n_{g'}}\right)}}$$

For the pymetrics data, the selection rates are: Asian (53.30%), Black (52.50%), Hispanic/Latino (56.80%), White (58.30%). The selection rates for other racial groups are lower than the White selection rate, but the disparities do not fall below the EEOC threshold since the impact ratios for all groups exceed 0.8. These results align with prior work from pymetrics [34]. However, pymetrics recommendations influence hiring at many distinct employers for many distinct positions. We argue that only analyzing impact ratios based on data aggregated from all pymetrics-mediated positions is an improper, or at minimum an incomplete, interpretation of the EEOC guidance, which was conceived for the purpose of identifying when a singular employer discriminates [13, 41 CFR 60-3.15.2(a)]. Instead, when applying this standard to a hiring algorithm vendor that mediates many hiring processes, each position should be analyzed *separately*. By analogy, if only men were recommended for doctor positions and only women were recommended for nurse positions, even if the selection rates matched so as to be acceptable in aggregate, the per-position discrepancies warrant scrutiny.

When we re-conduct our analysis on a per-position basis,[8] disaggregation reveals adverse impact against Black and Asian applicants.[9] 10.62% of positions adversely impact Black applicants and 30.70% of Black applicants apply to at least one of these positions.[10] 25.87% of all applications submitted by Black applicants are directed to these positions (39,986 applications). As some positions receive more applications than others, 23.3% of positions account for 80% of application-level adverse impact for this racial group. We see similar effects for Asians: 5.32% of positions demonstrate adverse impact against Asians, affecting 18.53% of Asian applicants and 14.74% of Asian applications (115,317 applications).[11] If applicants to the positions showing adverse impact were counterfactually recommended at the same rate as the most selected racial group, then an additional 11,513 Black applications and 29,320 Asian applications would have been recommended. Table 2 reports the aggregate position-level adverse impact by occupational group. Aggregating from individual positions to occupation groups suffices to mask the per-position adverse impact with the exception of Black and Asian applicants to Architecture and Engineering jobs.

We only attempt to identify adverse impact based on self-reported racial information and do not impute data, noting that we do not have access to features like applicant name. 62.35% of all applicants in the data do not

[7] All results for adverse impact exclusively consider applicants that self-report race.

[8] For our primary results, we apply the Benjamini-Hochberg correction, which bounds the false discovery rate at $\alpha = 0.05$, because we test multiple positions for adverse impact. However, the uncorrected results may be equally relevant in enforcing employment discrimination law since the EEOC may selectively investigate employers with positions that demonstrate adverse impact.

[9] Following EEOC guidance, results are for positions with at least 30 applicants who self-report race.

[10] Kline et al. [38] find that 7% of positions in their correspondence study discriminate against applicants with distinctively Black names.

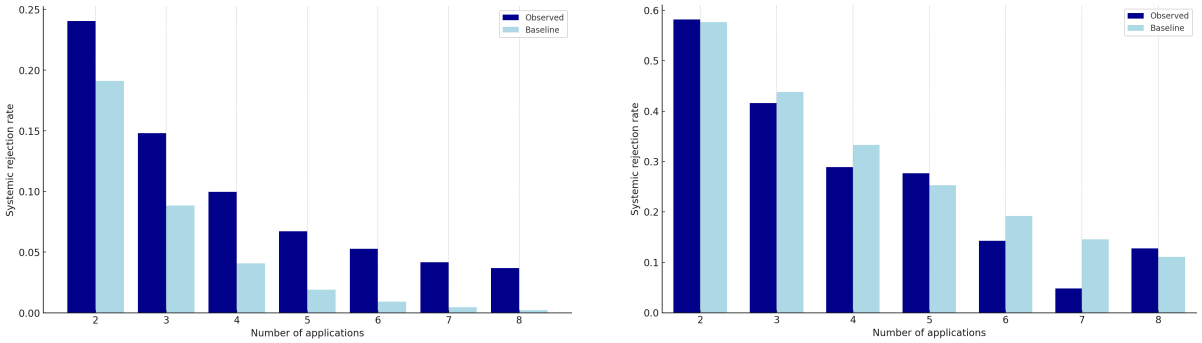
[11] pymetrics receives more Asian than Black applications, hence the discrepancy between relative and absolute.

self-report race as belonging to one of the four racial groups we study.[12] We expect our results underestimate the total amount of adverse impact as a result for two reasons: more applicants from adversely impacted groups apply to positions than we count and more positions demonstrate adverse impact than we count, because positions lacking at least 30 applicants with self-reported race are ruled out of the analysis.

3.2 Homogeneity in algorithmic hiring

Job seekers generally apply to multiple positions [16]. What happens when their applications are screened by algorithms from the same vendor? Prior work conjectures that an *algorithmic monoculture* [37] is likely to yield *homogeneous outcomes* [10]: some applicants are recommended for every job to which they apply and others are not recommended for any job. The latter case, *systemic rejection*, harms applicants because not being recommended by a first stage screening algorithm is very likely to result in rejection by the employer [20].

An applicant that is "algorithmically blackballed" [1] may be less likely to find a new job. This outcome is of special concern because a multidisciplinary literature establishes extended unemployment as harmful to individuals [39, 52, *inter alia*]. Extended unemployment may deplete financial resources and deteriorate both physical and mental health [51]. Repeated rejections may discourage future job pursuits and cause some applicants to leave the labor force altogether. If those that are systemically rejected are unable to contribute to the labor force, misallocation of talent may reduce cumulative economic production [25].



(a) Data from pymetrics

(b) Data from Kline et al. (2022) [38]

Fig. 2. **Systemic rejection rates.** The observed (dark blue) and baseline (light blue) systemic rejection rates for pymetrics data (left subplot) and a correspondence study of 108 U.S. firms [right subplot; 38].

When applying to two positions at two different employers, applicants might reasonably expect that they are receiving two separate evaluations and therefore two chances. But if both positions share the same model, their numerical score will be identical, which may violate applicants' expectations of different evaluations. We find that 42 models are shared across positions at different companies and there are 142 unique employer pairs that share at least one model.[13] Therefore, applicants in rare instances experience homogeneity mechanically due to *total algorithmic monoculture* [37].

[12]Note also that the categories pymetrics supplies for candidates to self-report their race reflect the U.S. Office of Management and Budget racial categories. The options may not include categories that the respondents believe best describe them or include the terms that best reflect racial or ethnic groups in their country or region. The candidates did have the option to select "Other".

[13]To identify shared models, we identify instances where the same model ID is used across different employers, though this may undercount model sharing if the same machine learning model is referred to using different model IDs.

While total algorithmic monoculture does occur, its impacts are limited in our dataset because very few applicants apply to positions at different employers served by the same underlying pymetrics model. Therefore, we study the more general and frequent form of algorithmic monoculture [10], where an applicant applies to multiple positions mediated by pymetrics models (Figure 2a), to understand how partial algorithmic monoculture relates to homogeneity. For example, applicants who apply to 4 positions can receive anywhere between 0 to 4 recommendations. 10% of these applicants are systemically rejected, as demonstrated by the observed (dark blue) bar at 4 applications. Since applicants are only permitted to take the pymetrics tests once every 330 days, similar weightings of a candidate's fixed gameplay features across different models will lead to similar outcomes for applicants across multiple applications within a year.

To contextualize the observed rates of systemic rejection, we distinguish systemic rejections rate foreseeable due to the underlying per-position rejection rates from further correlation between models. We introduce the *baseline* of independence as a tool to clarify the marginal increase in systemic rejection rate due to correlated model behavior [10, 57]. In §3.3, we support the validity of the baseline by showing that the largest empirical study of first-round screening procedures has a rate of systemic rejections consistent with the 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 applied to one position. In other words, compared to the observed systemic rate, which measures how often an applicant A is rejected from all three firms X, Y, and Z to which she applies, the baseline systemic rejection rate asks if three randomly selected applicants B, C, and D got rejected from X, Y, and Z, respectively, assuming the firms maintain their same rejection rates.

Formally, let N applicants each apply to k positions.[14] Define the outcome matrix $O \in \{0,1\}^{N\times k}$ such that $O[i,j]$ is the outcome for applicant i applying to position j, where 1 indicates recommendation and 0 indicates rejection. The selection rate for position j is $s_j = \frac{\sum_{i=1}^{N} O[i,j]}{N}$. For $t \in \{0,\ldots,k\}$, the observed rate at which applicants 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.

For all values of k, we find the observed systemic rejection rate significantly exceeds the baseline rate in the pymetrics data as shown in Figure 2a. A χ^2 goodness-of-fit test strongly rejects the null that observed systemic rejection rates match the baseline ($\chi^2 = 18{,}481$, $p < 0.001$). Our findings establish empirical evidence to support the claim that shared dependence on a single hiring algorithm vendor yields homogeneous outcomes.

3.3 Experimental Baseline

By contrast, we find that when first round screening is not mediated by a single screening procedure, systemic rejections are close to the baseline. To support the empirical validity of our baseline, we study homogeneous outcomes in the largest study of first-round screening at U.S. employers to date. Kline et al. [38] generated 83000 synthetic resumes and submitted these resumes to vacant positions at 108 US companies between October 2019 and April 2021, a similar time period to our data. The companies, which are a subset of the Fortune 500,[15] collectively employ 15 million workers. We analyze the homogeneity observed in the resulting callback outcomes in their data.

[14] The above notation (for simplicity) assumes all applicants apply to the same k positions: the general form is Appendix C.

[15] 10 of the companies are parent companies of Fortune 500 companies.

We find that the baseline *is* an effective estimator of the systemic rejection rate for this dataset. As shown in Figure 3, the observed systemic rejection rate is accurately predicted by the baseline and a chi-squared goodness-of-fit test cannot reject equality of the two distributions ($\chi^2 = 20.05$, $p = 0.69$). In other words, while the largest previous study observes systemic rejection rates consistent with employers making statistically independent decisions, the algorithmic hiring data shows significantly correlated outcomes that lead to higher-than-baseline systemic rejection rates. The full results are shown in the appendix in Table 6.

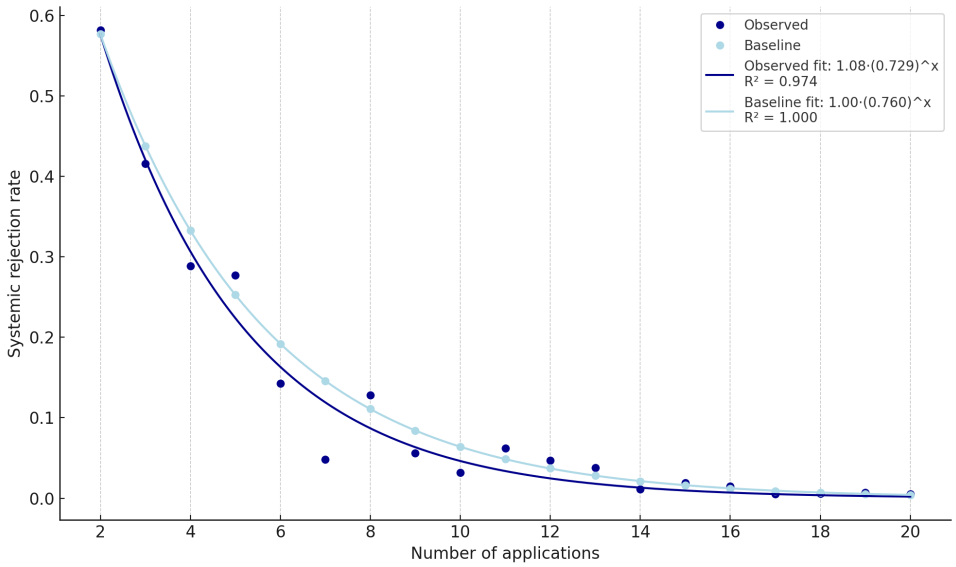


Fig. 3. **Kline et al. (2022) systemic rejection rates.** We plot the observed (dark blue) and systemic rejection rates (light blue) for data from a correspondence study of 108 U.S. employers [38] along with the associated exponential fits.

In order to perform this analysis, we needed to decide which resumes should be considered the same. Due to [38]'s study design, the exact same resume is generally not submitted to multiple positions. Since we are interested in the outcomes an applicant receives when they apply to multiple positions, we treat two resumes as belonging to the same applicant if they have the same values for the following variables: firstname, lastname, race, gender, over40, associates, lgbtq_club, political_club, academic_club, gender_neutral_pronouns, same_gender_pronouns. These variables are the targets of interest for the study. Other variables are chosen to make the resumes equally appealing to each employer. In particular, applicant addresses are chosen such that they are close to the location of the posted job. Therefore we consider a resume to be the same if it has the features listed above but a different home address. Under this assumption, we report the observed and baseline systemic rejection rates in Table 6 in the appendix.

While both datasets involve large US employers during 2019–2020 and include entry-level positions, they differ in several ways. Kline et al. [38] study callbacks—employer-initiated contact attempts—occurring at rates of 23–25%, whereas pymetrics assessments yield pass rates of approximately 50%. Their sample is restricted to entry-level positions, whereas our data span entry-level through director roles. The geographic scope also differs: Kline et al. [38] construct applicant profiles using addresses from public high schools within the county of each job posting, whereas our pymetrics data reflect global hiring—6.56% of applicants report India as their country of residence, and the most common city in our data is London. Although some Kline et al. [38] employers

used personality tests, pymetrics' game-based assessments would be difficult to manipulate similarly, suggesting minimal vendor overlap. Algorithmic screening is universal in our data but likely variable in theirs: some employers used it, others did not, and those that did use it likely relied on different vendors.

Since our homogeneity measure adjusts for the underlying positive rate, differences in callback versus pass rates do not mechanically explain the divergent findings. Nevertheless, the baseline closely tracks observed homogeneity in Kline et al. [38]'s data but not in our data. This suggests that the excess homogeneity we document may be a distinctive feature of centralized algorithmic assessment systems, in which models from a single provider evaluate all applicants, rather than a general property of candidate screening processes.

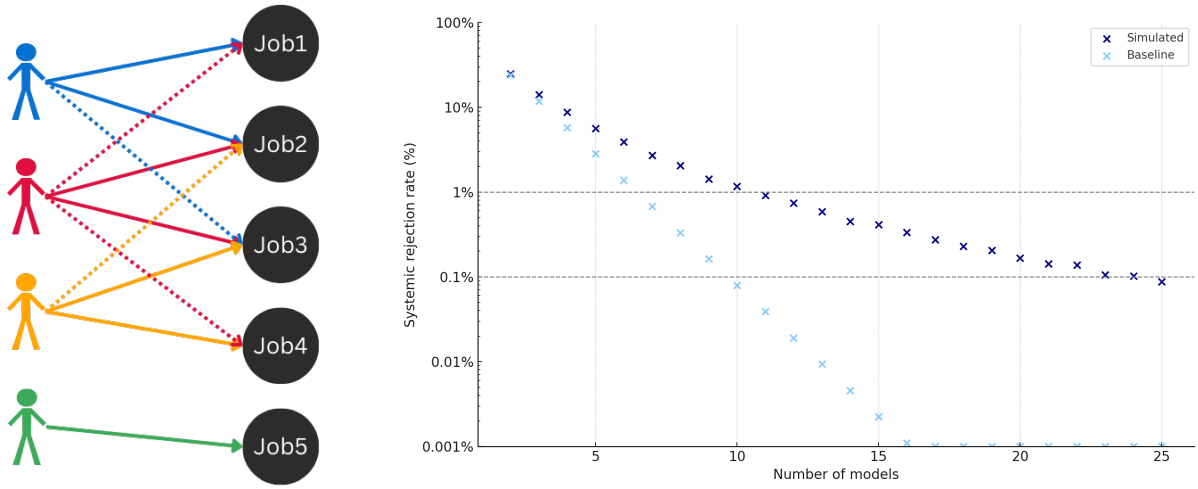


Fig. 4. **Simulation results. Left:** Given the observed set S of models an applicant is assessed by (solid lines), the *connected set* $S' \supseteq S$ includes every model that assessed an applicant assessed by a model in S (solid and dashed lines). **Right:** Systemic rejection rate as a function of number of models sampled from S'.

3.4 Large-scale simulation of algorithmic hiring outcomes

Our empirical findings show that applicants face both adverse impact and systemic rejection. Could applicants avoid systemic rejection by applying to more positions or to different positions? Studying counterfactual outcomes is challenging for conventional methods in labor economics that rely on observational data. In the real world, an applicant only receives an outcome if they apply to a position. However, we leverage the deterministic replicability of pymetrics hiring algorithms and the efficiency of algorithmic decision-making to simulate the outcomes applicants would have received if they applied to every position mediated by pymetrics algorithms. We sample 1000 applicants at random and ask pymetrics to evaluate them against each of the applicable 495 models.[16] Let $O_{\text{sim}} \in \{0,1\}^{939\times 495}$ denote the simulated binary outcome matrix where $O[i,j]$ indicates if applicant i was recommended by model j. We find that no sampled applicant is rejected by every pymetrics model. The applicant who receives the fewest recommendations is still recommended by 52 models (11%).

However, while most applicants apply to multiple positions, no applicants apply to *all* positions in the labor market. Applicants are less likely to apply to jobs in distant places, in different market sectors, or during times

[16]Since these models were developed over several years by pymetrics, during which time they changed their infrastructure, only 939 applicants have simulated outcomes.

when they are happily employed [6, 41, 43]. Since it is unrealistic for applicants to apply to every position, we refine our analysis by studying an intermediate regime. In this intermediate regime, applicants (i) apply more broadly than they did in reality yet (ii) do not apply to every position. For an applicant, given the observed set S of models they applied to, we define their *connected set* $S' \supseteq S$ as models that share applicants with the models in S (left of Figure 4) because applicants might share preferences or constraints with applicants who applied to jobs to which the applicant also applied. For the sampled applicants, the average set size grows from 1.23 observed models to 145 connected models.

Formally, let $A \in \{0,1\}^{N \times M}$ be the observed application matrix where N is the total number of applicants (i.e. 3,372,132), M is the total number of models (i.e. 555) and $A[i,j]$ indicates if applicant i was assessed by model j in reality. $B = A^T A$ is the model-model overlap matrix that encodes the number of applicants shared by every pair of models in the observed data. Let $A_{\text{sim}} \in \{0,1\}^{939 \times 495}$ be the submatrix of the observed application matrix A corresponding to the applicants and models we simulate outcomes for. Similarly, let $B_{\text{sim}} \in \mathbb{R}^{495 \times 495}$ be the submatrix of the model-model overlap matrix B corresponding to the models we simulate outcomes for. $A'_{\text{sim}} = \min(1, A_{\text{sim}} B_{\text{sim}}) \in \{0,1\}^{939 \times 495}$ is the application matrix corresponding to our intermediate regime where applicants apply (i) to all models they applied to in reality and (ii) all models that share an applicant with those they applied to.[17] A'_{sim} expands the observed application structure yet sparsifies the fully dense structure:

$$A_{\text{sim}} \leq A'_{\text{sim}} \leq \mathbf{1}_{939 \times 495}$$

In our analyses involving homogeneous outcomes, we prefer to study settings that fix the number of applications k submitted by each applicant. To achieve this level of control, we sample applicant outcomes so every considered applicant submits k applications. Namely, for each applicant i, we discard the corresponding row in A'_{sim} if they applied to fewer than k models (i.e. $\sum A'_{\text{sim}}[i] < k$) and we sample k models if they applied to at least k models.

In Figure 4 (right), we vary the number of sampled models from each applicant's connected set of models. Both the simulated and baseline data are well-described by exponential functions as shown by their linear scaling on the plot with a logarithmically-spaced y-axis. The systemic rejection falls below 0.1% first at 25 models for simulation, compared to 10 under the baseline. That is, under the proposed applicant behavior, applicants need to apply to at least 25 different positions to ensure at least one recommendation with high probability (i.e. 99.9%). Given that the same model might be used at different positions, the number could be higher in practice. Since a pymetrics recommendation only admits an applicant to the pool of applications considered by a human recruiter, applicants would need to apply to even more jobs in order to increase their likelihood of an interview.

4 Discussion

Hiring algorithms change lives and shape labor markets, but we lack empirical research into their impact on applicants. Most prior work centers the perspective of employers, measuring whether an algorithm reflects employer preferences, reduces employer costs, and complies with employer-level employment discrimination law. While important, the employer-centric perspective neglects the structural shifts to the labor market caused by algorithms. Algorithms not only influence hiring at each employer but also link outcomes *across* employers due to shared dependence on the same vendor(s). Since an applicant's employment status depends on the *cumulative decisions* of many employers, algorithmic monoculture may precipitate systemic exclusion from the labor market for some applicants. We discuss the relationship between our work and policy (§4.1) as well as key limitations relevant to the interpretation of this work and the prospect of future work (§4.2).

[17] We describe the matrix in this way for clarity. However, condition (ii) suffices to exactly characterize the matrix, because a model must share an applicant with itself, namely the applicant being considered.

4.1 Policy

Three types of policy bear on algorithmic hiring: (i) policies that govern hiring, (ii) policies that govern AI/algorithmic decisions and (iii) policies that specifically govern algorithmic hiring. In the first category, the U.S. federal law that pertains to our research on discrimination is Title VII of the Civil Rights Act of 1964. This is the origin of the impact ratio standard, or the "4/5ths rule", described in this paper. Notably, the theory of disparate-impact liability, including in the Title VII context, is subject to scrutiny under the second Trump administration's Executive Order 14281 from 2025 entitled "Restoring Equality of Opportunity and Meritocracy". At the time of writing, it remains unclear how this scrutiny will affect future US employment discrimination policy.

In the second category, we highlight the European Union's AI Act and, specifically, the regulation of high-risk AI systems. Notably, Annex III of the E.U. AI Act provides a default high-risk designation for AI systems on the E.U. market that relate to "employment, workers' management and access to self-employment".[18] The systems we study likely qualify as high-risk under this designation and as of August 2, 2026, system providers and deployers will be subject to several compliance requirements. While our measurement of adverse impact and the racial categories offered for candidate self-report by pymetrics are both tied to U.S. law, our methods may be informative for risk management for other high-risk hiring AI systems under the E.U. AI Act.

In the third category, we highlight New York City Local Law 144, which set precedent on directly regulating algorithmic hiring. However, since its passing in 2021, research demonstrates its limited efficacy due to issues of null compliance where employers exercise discretion over whether they are in scope of the law [61]. In light of our findings, we stress that existing government guidance for Local Law 144 does not address the distinction we raise between aggregate and position-level impact ratios.

We hope that our research can contribute to future evidence-based AI policy [9] and therefore offer the following recommendations based on our findings.[19]

Recommendation 1: Regulators and auditors should measure adverse impact per position. We find significant adverse impact at the position level that is masked in aggregate. Since employment standards already consider position-level adverse impact, these standards should apply to measurement in the algorithmic setting, appropriately adapting to whether the data describes applications to a single position, a single employer with multiple positions, or multiple employers and positions. Our recommendation may contradict existing guidance for New York City Local Law 144 of 2021: "The vendor provides historical data regarding applicant selection that the vendor has collected from multiple employers to an independent auditor who will conduct a bias audit as follows:". The government-provided example computation of impact ratios appears to indicate all the data should be merged together, blurring distinctions between positions and even employers.[20]

Recommendation 2: Agencies should strengthen market surveillance. Existing authority enables agencies to partially understand employer-level hiring practices. For example, the EEOC collects annual EEO-1 reports on employee demographics for companies with at least 100 employees. Since employment is essential for individual welfare [2, 59], there is a strong social imperative to understand homogeneous outcomes that precipitate systemic exclusion from the labor market. If systemic rejection mediated by algorithmic hiring leads an applicant to be unemployed, or unemployed for significantly longer than they would otherwise be, the length of unemployment can be a cascading harm, decreasing the likelihood of future interviews [19, 40]. Current application of agency

[18]The full text identifies two categories: (a) AI systems intended to be used for the recruitment or selection of natural persons, in particular to place targeted job advertisements, to analyse and filter job applications, and to evaluate candidates and (b) AI systems intended to be used to make decisions affecting terms of work-related relationships, the promotion or termination of work-related contractual relationships, to allocate tasks based on individual behaviour or personal traits or characteristics or to monitor and evaluate the performance and behaviour of persons in such relationships.

[19]For specificity, we refer to the U.S. context due to greater relevance of our findings and familiarity with this policy environment, but they may generalize to other jurisdictions.

[20]See https://codelibrary.amlegal.com/codes/newyorkcity/latest/NYCrules/0-0-0-138393.

authority is ill-equipped to address the homogeneous outcomes we find for three reasons. First, not all systemic rejections will align with attributes that agencies are authorized to investigate because they are protected under discrimination law, such as race, age, or gender. Second, existing data collection (e.g. EEO-1) is aggregated and anonymized, precluding linking of outcomes across employers. Where legitimate privacy interests prevent even anonymized linking of outcomes, agencies should explore alternatives, perhaps through monitors for the unemployment rate and length of job searches such as the Current Population Survey (CPS). Alternatively, akin to our work, agencies should explore applying their investigative authorities to acquire access to unique centralized datasets such as those owned by hiring vendors, which are likely to have the outcomes for the same applicants across many applications. Third, existing techniques to measure market concentration, while valuable, will not be an accurate measure of systemic rejection. Models created by the same company might be more or less diverse in their rejections depending on modeling techniques, and models created by different companies can reject the same individuals, showing less diversity than might have been expected [30], especially when they rely on "shared components" such as training data, model architecture, or base model that correlate their outcomes [10, 36].

Recommendation 3: Agencies should monitor algorithmic monoculture. Our work demonstrates specific negative outcomes, namely racial adverse impact and systemic rejection, under the conditions of algorithmic monoculture. Alongside measuring these risks, agencies should also consider the potential underlying structural cause. How prevalent is algorithmic monoculture in hiring? Awareness could anticipate harm, even if current recourse is primarily reactive, such as litigation in response to discrimination.

In addition to their potential contribution to adverse impact and systemic exclusion, the phenomena we study in this work, algorithmic hiring monocultures may be of interest for other reasons. As with other forms of algorithmic monoculture, system-wide resilience may be compromised if hiring algorithm vendors experience outages: for example, if HireVue was unable to provide algorithmic recommendations for an extended period, hiring may be delayed or disrupted across thousands of employers including federal agencies [44]. Further, other domains of employment regulation require that competitors maintain separate practices (e.g. employers cannot collude to set wages). Therefore, if competitors depend on the same hiring algorithm vendor to inform their hiring decisions, possibly by pooling data to train the vendor's algorithms, competition in hiring may be reduced to the disadvantage of applicants, as algorithmic monoculture has been shown to do in other domains [31]. As agencies monitor how algorithmic dependence manifests in labor markets, policymakers should consider what levels of entanglement are (un)acceptable.

Recommendation 4: Legislators should consider whether to mandate researcher access to algorithmic hiring. Our work demonstrates the value of independent research, especially given the dearth of empirical research on hiring algorithms. Empirical methods for studying hiring such as correspondence studies and surveys are effective for studying employment outcomes as well as the experiences of both applicants and employers, but struggle to isolate the effects of intermediary components such as hiring algorithms. [21] Most prior work on algorithmic hiring relies on scraping data from publicly accessible job posting platforms like Indeed and LinkedIn [63]. Research on hiring algorithms is stymied by researchers' inability to access hiring algorithms and their predictions.

pymetrics is a noteworthy exception that has published research [34] and provided data to external paid research collaborators [60]. However, we expect further empirical progress on hiring algorithms will be hard to come by under current conditions. Social media researchers historically faced similar challenges: "platforms likely will not make data available without binding legal mechanisms" yet "access to data for empirical research is seen as a necessary step in ensuring transparency and accountability" [45]. In response, the European Union's Digital Services Act of 2022 requires very large online platforms (i.e. platforms with at least 45 million EU users) to provide researcher access since these platforms wield substantial and increasingly monopolistic control over

[21] Correspondence studies have also been subject to critique on the grounds of the use of deception and measurement validity [22, 35].

information online. Legislators should contemplate similar policy given (i) the scale of algorithmic adoption in hiring, (ii) the stakes of hiring decisions, and (iii) the absence of compelling alternatives for data-driven inquiry.

4.2 Limitations

The core limitations we identify pertain to (i) generalizability to algorithmic hiring as a whole, (ii) relationships with notions of applicant quality and model validity, (iii) relevance to the enforcement of US employment law, and (iv) certainty about downstream hiring outcomes. In addition, we acknowledge that our work may suppress future empirical research into algorithmic hiring and may therefore prompt stronger forms of policy response to ensure researcher access to data.

Our findings may not generalize to all algorithmic screening. We do not have data from other vendors nor are aware of viable mechanisms to acquire such data. In particular, the game-based approach of pymetrics may qualitatively differ from alternative approaches such as resume screening.

We lack external measures of both applicant quality and model validity. Because we cannot measure applicant quality, we cannot predict if applicants that were systemically rejected would have been effective at the positions to which they applied. Lack of independent information about applicant quality affects our finding of homogenization but not our finding of adverse impact. No demonstration of applicant quality or fit to job is necessary to begin an investigation of employment discrimination; establishing adverse impact can be sufficient. We also lack external measures of model validity. We are therefore uncertain of how accurately pymetrics models reflect the preferences of the associated employers or how successful the models are at selecting candidates who are good fits for the positions.

We cannot determine whether the evidence of adverse impact we present implies the existence of a "less discriminatory" algorithm or procedure that serves the same business need [8, 78-9]. Under the relevant US law for Title VII, while evidence of adverse impact may suffice to initiate an investigation and provide probative value in litigation, it does not necessarily imply an employer's conduct is illegal. In particular, we do not have visibility into the development process for pymetrics models to understand whether alternative models could have been selected or generated at similar cost. The closest comparison we do make is to a large-scale correspondence study data [38], which does not represent a single alternative screening procedure, as the acceptances and rejections were performed by a mix of human and algorithmic decision-making at many different firms.

We do not know how the pymetrics recommendations are used by hiring managers at employers to make final hiring decisions. Some prior work suggests that following the recommendations of hiring algorithms leads to identifying non-traditional applicants and/or better applicants [14, 23]; others finds that human judgments mediated by algorithmic recommendations are more discriminatory than human judgments alone [11]. Final human decisions may be influenced by a suite of algorithmic recommendations, and different companies likely use different additional algorithmic tools in their screening processes [53]. However, we believe that not being recommended by a pymetrics model means the downstream employer is very likely to reject the applicant, which partially abates this uncertainty in the context of (systemic) rejections.

Finally, we hope that this work encourages further independent research into algorithmic hiring, including transparency into and scrutiny of algorithmic hiring vendors. However, our work in itself may discourage the voluntary data sharing and demographic data collection that made these findings possible. Therefore, our work may need to serve as a foundation for policy that enables deeper inquiry into major hiring algorithm vendors. This bears a close resemblance to the context around social media and online platform research that prompted mandatory researcher data access under the European Union's Digital Services Act [45].

5 Generative AI Usage Statement

The authors have used Generative AI tools to generate code used to generate tables, plots and macros to input the numerical values (Claude Opus 4); some authors have used it to edit portions of the text (Claude Sonnet 4).

Acknowledgments

We thank Alondra Nelson, Anna Stansbury, Arvind Narayanan, Ashia Wilson, Bo Cowgill, Christo Wilson, Dan Ho, Deb Raji, Emma Pierson, Erik Brynjolfsson, Ifeoma Ajunwa, Jon Kleinberg, Judy Shen, Kathleen P. Nichols, Kristian Lum, Lindsey Raymond, Manish Raghavan, Meena Jagadeesan, Mina Lee, Omer Reingold, Reva Schwartz, Rob Reich, Roger Creel, Sanmi Koyejo, Sayash Kapoor, Shibani Santurakar, Shomik Jain, Solon Barocas, Sonny Tambe, Suresh Venkatasubramanian, and Zachary Bleemer for their thoughtful feedback and guidance. We also thank Leo He, Sharmeen Malik, Frida Polli, Shea Valentine, and Georgiy Yudintsev for their support and help with accessing and understanding the data.

References

[1] Ifeoma Ajunwa. 2021. An Auditing Imperative for Automated Hiring. *Harvard Journal of Law and Technology* 1 (2021). Issue 34.

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

[3] David Autor and David Scarborough. 2008. Does Job Testing Harm Minority Workers? Evidence from Retail Establishments. *The Quarterly Journal of Economics* (February 2008).

[4] Solon Barocas and Andrew Selbst. 2016. Big Data's Disparate Impact. 104 (2016), 671–732. https://doi.org/10.15779/Z38BG31

[5] BBC. 2024. AI hiring tools may be filtering out the best job applicants. (2024). https://www.bbc.com/worklife/article/20240214-ai-recruiting-hiring-software-bias-discrimination

[6] Michèle Belot, Philipp Kircher, and Paul Muller. 2018. Providing Advice to Jobseekers at Low Cost: An Experimental Study on Online Advice. *The Review of Economic Studies* 86, 4 (10 2018), 1411–1447. https://doi.org/10.1093/restud/rdy059 arXiv:https://academic.oup.com/restud/article-pdf/86/4/1411/28883098/rdy059.pdf

[7] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (Aug. 2004), 991–1013. https://doi.org/10.1257/0002828042002561

[8] Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. 2024. Less Discriminatory Algorithms. *Georgetown Law Journal* 113, 1 (2024).

[9] Rishi Bommasani, Sanjeev Arora, Jennifer Chayes, Yejin Choi, Mariano-Florentino Cuéllar, Li Fei-Fei, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Scott Singer, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, and Dawn Song. 2025. Advancing science- and evidence-based AI policy. *Science* 389, 6759 (2025), 459–461. https://doi.org/10.1126/science.adu8449 arXiv:https://www.science.org/doi/pdf/10.1126/science.adu8449

[10] Rishi Bommasani, Kathleen Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. Picking on the Same Person: Does Algorithmic Monoculture Homogenize Outcomes?. In *Advances in Neural Information Processing Systems.*

[11] Moa Bursell and Lambros Roumbanis. 2024. After the algorithms: A study of meta-algorithmic judgments and diversity in the hiring process at a large multisite company. *Big Data amp; Society* 11, 1 (Jan. 2024). https://doi.org/10.1177/20539517231221758

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

[13] Equal Employment Opportunity Commission. 1978. Uniform Guidelines on Employee Selection Procedures. , 38295-38314 pages. https://www.ecfr.gov/current/title-41/section-60-3.15

[14] Bo Cowgill. 2020. Bias and Productivity in Humans and Algorithms: Theory and Evidence from Résumé Screening. (21 March 2020). https://conference.iza.org/conference_files/MacroEcon_2017/cowgill_b8981.pdf Presented at IZA Workshop: Labor Productivity and the Digital Economy, OECD, Paris, October 30-31, 2017.

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

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

[17] U.S. EEOC. 1979. Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures. *Federal Register* 44, 43 (2 March 1979). https://web.archive.org/web/20250524183445/https://www.eeoc.gov/laws/guidance/questions-and-answers-clarify-and-provide-common-interpretation-uniform-guidelines Title VII, 29 CFR Part 1607.

[18] Alessandro Fabris, Clara Rus, Jorge Saldivar, Anna Gatzioura, Asia J. Biega, and Carlos Castillo. 2026. Does fair ranking lead to fair recruitment outcomes? A study of interventions, interfaces, and interactions. *Information Processing Management* 63, 3 (2026), 104506. https://doi.org/10.1016/j.ipm.2025.104506

[19] Henry S. Farber, Chris M. Herbst, Dan Silverman, and Till von Wachter. 2019. Whom Do Employers Want? The Role of Recent Employment and Unemployment Status and Age. *Journal of Labor Economics* 37, 2 (2019), 323–349. https://doi.org/10.1086/700184 arXiv:https://doi.org/10.1086/700184

[20] J. Fuller, M. Raman, E. Sage-Gavin, and K. Hines. 2021. *Hidden Workers: Untapped Talent.* Technical Report. Harvard Business School Project on Managing the Future of Work and Accenture.

[21] S. Michael Gaddis. 2014. Discrimination in the Credential Society: An Audit Study of Race and College Selectivity in the Labor Market. *Social Forces* 93, 4 (Nov. 2014), 1451–1479. https://doi.org/10.1093/sf/sou111

[22] Daniel Hamermesh. 2012. Are Fake Resumes Ethical for Academic Research? Freakonomics Blog. http://freakonomics.com/2012/03/15/are-fake-resumes-ethical-for-academic-research/ Accessed: 2025-04-22.

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

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

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

[26] Lily Hu. 2023. What is "Race"' in Algorithmic Discrimination on the Basis of Race? *Journal of Moral Philosophy* 21, 1-2 (2023), 1 – 26. https://doi.org/10.1163/17455243-20234369

[27] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. ACM, 375–385. https://doi.org/10.1145/3442188.3445901

[28] Louis S. Jacobson, Robert J. LaLonde, and Daniel G. Sullivan. 1993. Earnings Losses of Displaced Workers. *The American Economic Review* 83, 4 (1993), 685–709. http://www.jstor.org/stable/2117574

[29] Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. 2024. Algorithmic Pluralism: A Structural Approach To Equal Opportunity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM. https://doi.org/10.1145/3630106.3658899

[30] Shomik Jain, Margaret Wang, Kathleen Creel, and Ashia Wilson. 2025. Allocation Multiplicity: Evaluating the Promises of the Rashomon Set. https://doi.org/10.48550/ARXIV.2503.16621

[31] Nathanael Jo, Kathleen Creel, Ashia Wilson, and Manish Raghavan. 2025. Homogeneous Algorithms Can Reduce Competition in Personalized Pricing. *Neural Information Processing Systems (NeurIPS)* (2025). https://arxiv.org/abs/2503.15634

[32] Gabbrielle M. Johnson. 2020. Algorithmic bias: on the implicit biases of social technology. *Synthese* 198, 10 (2020), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y

[33] Gabbrielle M. Johnson. 2025. The hard proxy problem: proxies aren't intentional; they're intentional. *Philosophical Studies* 182, 5–6 (May 2025), 1383–1411. https://doi.org/10.1007/s11098-025-02333-9

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

[35] Judd B. Kessler, Corinne Low, and Colin D. Sullivan. 2019. Incentivized Resume Rating: Eliciting Employer Preferences without Deception. *American Economic Review* 109, 11 (November 2019), 3713–44. https://doi.org/10.1257/aer.20181714

[36] Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. Correlated Errors in Large Language Models. arXiv:2506.07962 [cs.CL] https://arxiv.org/abs/2506.07962

[37] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021). https://doi.org/10.1073/pnas.2018340118 arXiv:https://www.pnas.org/content/118/22/e2018340118.full.pdf

[38] Patrick Kline, Evan K Rose, and Christopher R Walters. 2022. Systemic Discrimination Among Large U.S. Employers*. *The Quarterly Journal of Economics* 137, 4 (06 2022), 1963–2036. https://doi.org/10.1093/qje/qjac024 arXiv:https://academic.oup.com/qje/article-pdf/137/4/1963/51053979/qjac024.pdf

[39] Kory Kroft, Fabian Lange, and Matthew J. Notowidigdo. 2013. Duration Dependence and Labor Market Conditions: Evidence from a Field Experiment*. *The Quarterly Journal of Economics* 128, 3 (06 2013), 1123–1167. https://doi.org/10.1093/qje/qjt015 arXiv:https://academic.oup.com/qje/article-pdf/128/3/1123/30631486/qjt015.pdf

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

[41] Peter Kuhn and Kailing Shen. 2023. What Happens When Employers Can No Longer Discriminate in Job Ads? *American Economic Review* 113, 4 (April 2023), 1013–48. https://doi.org/10.1257/aer.20211127

[42] Peter Kuhn, Kailing Shen, and Shuo Zhang. 2020. Gender-targeted job ads in the recruitment process: Facts from a Chinese job board. *Journal of Development Economics* 147 (Nov. 2020), 102531. https://doi.org/10.1016/j.jdeveco.2020.102531

[43] Ioana Marinescu and Roland Rathelot. 2018. Mismatch Unemployment and the Geography of Job Search. *American Economic Journal: Macroeconomics* 10, 3 (July 2018), 42–70. https://doi.org/10.1257/mac.20160312

[44] Allie Nawrat. 2023. Inside HireVue's acquisition of Modern Hire. https://www.unleash.ai/hr-technology/inside-hirevues-acquisition-of-modern-hire/ Accessed on March 25, 2024.

[45] Brandie Nonnecke and Camille Carlton. 2022. EU and US legislation seek to open up digital platform data. *Science* 375, 6581 (2022), 610–612. https://doi.org/10.1126/science.abl8537 arXiv:https://www.science.org/doi/pdf/10.1126/science.abl8537

[46] pymetrics. 2020. audit-AI: How we use it and what it does. https://web.archive.org/web/20250305042721/https://github.com/pymetrics/audit-ai/blob/master/examples/implementation_suggestions.md

[47] Lincoln Quillian and John J. Lee. 2023. Trends in racial and ethnic discrimination in hiring in six Western countries. *Proceedings of the National Academy of Sciences* 120, 6 (Jan. 2023). https://doi.org/10.1073/pnas.2212875120

[48] Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen. 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences* 114, 41 (2017), 10870–10875. https://doi.org/10.1073/pnas.1706255114 arXiv:https://www.pnas.org/content/114/41/10870.full.pdf

[49] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 469–481. https://doi.org/10.1145/3351095.3372828

[50] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. 2022. An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Mining and Knowledge Discovery* 36, 6 (Sept. 2022), 2153–2193. https://doi.org/10.1007/s10618-022-00861-0

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

[52] Amartya Sen. 1997. Inequality, unemployment and contemporary Europe. *Int'l Lab. Rev.* 136 (1997), 155.

[53] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. 2022. A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns* 3, 2 (Feb. 2022), 100425. https://doi.org/10.1016/j.patter.2021.100425

[54] Mona Sloane, Ian René Solano-Kamaiko, Jun Yuan, Aritra Dasgupta, and Julia Stoyanovich. 2023. Introducing contextual transparency for automated decision systems. *Nature Machine Intelligence* 5, 3 (March 2023), 187–195. https://doi.org/10.1038/s42256-023-00623-7

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

[56] Lex Thijssen, Frank van Tubergen, Marcel Coenders, Robert Hellpap, and Suzanne Jak. 2021. Discrimination of Black and Muslim Minority Groups in Western Societies: Evidence From a Meta-Analysis of Field Experiments. *International Migration Review* 56, 3 (Nov. 2021), 843–880. https://doi.org/10.1177/01979183211045044

[57] Connor Toups, Rishi Bommasani, Kathleen Creel, Sarah Bana, Dan Jurafsky, and Percy S Liang. 2023. Ecosystem-level Analysis of Deployed Machine Learning Reveals Homogeneous Outcomes. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 51178–51201. https://proceedings.neurips.cc/paper_files/paper/2023/file/a0b1082fc7823c4c68abcab4fa850e9c-Paper-Conference.pdf

[58] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. 2021. When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring1. *Management Information Systems Quarterly* 45, 3 (09 2021), 1557–1580. https://doi.org/10.25300/MISQ/2021/16559

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

[60] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 666–677. https://doi.org/10.1145/3442188.3445928

[61] Lucas Wright, Roxana Mika Muenster, Briana Vecchione, Tianyao Qu, Pika (Senhuang) Cai, Alan Smith, Comm 2450 Student Investigators, Jacob Metcalf, and J. Nathan Matias. 2024. Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1701–1713. https://doi.org/10.1145/3630106.3658998

[62] Meg Young, Michael Katell, and P.M. Krafft. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1375–1386. https://doi.org/10.1145/3531146.3533194

[63] Shuo Zhang and Peter Kuhn. 2022. Understanding Algorithmic Bias in Job Recommender Systems: An Audit Study Approach. (2022).

A Data

We acquire large-scale data from pymetrics. To better contextualize this data, we describe the terms that govern its use and provide greater detail into the data and its processing.

A.1 Data terms

We signed a Data Use Agreement (DUA) with pymetrics that grants access to the data through December 31, 2025. The DUA imposes no restrictions on publication beyond a 30-day review period, during which pymetrics may request removal of Company Confidential Information only. We received the data from pymetrics through a series of secure data transfers subject to our data use agreement with pymetrics and our data risk assessment with our research institution. The DUA also restricts redistribution of data by the researchers. The fully executed DUA is available at the conclusion of this appendix (with non-substantive redactions to preserve anonymity during the review process). The research was subject to IRB review and data risk assessment by the research institution.

Data Access Agreement

This Data Access Agreement ("Agreement") is between [University] ("University"), an institution of higher education, and pymetrics Inc. ("Company"), a corporation, is effective on the [date] ("Effective Date").

Company plans to provide data described as "A database, including the following fields: anonymized candidate IDs, scores indicating candidate fit to range of pymetrics models, corresponding ONET codes for pymetrics models, demographic data collected from candidates via pymetrics' exit screen" ("Data") to [Principal Investigator] ("Principal Investigator"), who is a University employee, for a research project set forth in Exhibit A ("Research Program"). The parties hereby agree as follows:

GRANT AND TRANSFER

1.1 **Grant.** Subject to the terms and conditions of this Agreement, Company grants University the nonexclusive right to use the Data solely in the Research Program.

1.2 **Transfer Term.** Company will make the Data available to University during the term of this Agreement, a period from: [start date] to [end date] ("Term"). The Term may be extended only by advance written agreement of both parties.

1.3 **No Other Rights.** This Agreement does not constitute, grant nor confer any license under any patents or other proprietary interests of one party to the other, except as explicitly stated in this Agreement.

1.4 Each party shall retain all right, title, and interest in its respective technology and intellectual property first conceived or reduced to practice or fixed in a tangible medium by such party before the Effective Date and independent of such party's performance of this Agreement, together with all intellectual property rights in or to the foregoing ("Background Intellectual Property"). No right, title, or interest to either party's Background Intellectual Property shall transfer to the other party under this Agreement.

1.5 Company will own all right, title and interest to intellectual property first conceived and reduced to practice and/or fixed in a tangible medium solely by Company's personnel that relates to Company's Data, technical information, trade secrets, know-how and any source code disclosed by Company ("Proprietary Information") or Company's Background Intellectual Property, together with any improvements, modifications or derivative works thereof made by either party either solely or jointly (and all intellectual property rights in or to the foregoing) ("Company Intellectual Property"). University and Company will jointly own intellectual property first conceived and reduced to practice and/or fixed in a tangible medium jointly by University's and Company's personnel during the Term and directly arising from the Research Program, excluding improvements, modifications or derivative works of Proprietary Information, Company Intellectual Property, or Company Background Intellectual Property (and all intellectual property rights in or to the foregoing) ("Joint Intellectual Property"). University will own all right, title and interest to intellectual property first conceived and reduced to practice and/or fixed in a tangible medium solely by University's personnel during the term of and directly arising from the Research Project, excluding Proprietary Information, Company Background Intellectual Property, or Company Intellectual Property (or improvements, modifications or derivative works thereof and all intellectual property rights in or to the foregoing) included or incorporated therein ("Project Intellectual Property"). University hereby grants to Company a non-exclusive, royalty-free license to Project Intellectual Property, and University's rights, title, and interest in Joint Intellectual Property solely for Company's internal research and development. University will execute any documents reasonably requested by Company to effect or perfect all rights, title, and interest of Company under this section and this Agreement. Notwithstanding the foregoing, for the sake of clarity, both University and Company agree that no Joint Intellectual Property is foreseen to result from this Agreement. If there is any Joint Intellectual Property that is developed under this Agreement, Company and University will negotiate in good faith to find mutual agreement about the disposition of such Joint Intellectual Property.

COMPANY DATA

2.1 **Ownership.** Company retains ownership of Data. Company retains all rights to distribute the Data to other commercial or non-commercial entities. Before University's use, University shall ensure that Data shall be de-identified and shall not be used for any other purpose than those contemplated herein. University will not knowingly take any action that will enable re-identification of any Data subject.

2.2 **Authority.** Company warrants it has the authority to provide Data to University for use in the Research Program.

UNIVERSITY USE OF COMPANY DATA

3.1 **Restrictions.** University will use Data only for the Research Program as specified herein. If University seeks to use Data for other purposes, University will obtain written consent from Company, either by an amendment to this Agreement or a new agreement, before such use. University shall (1) store all Data received under this Agreement using encrypted and password protected devices; (2) maintain a list of all individuals with access to Data, as well as a log of such access, (3) restrict access to team members on a need to know basis and (4) destroy Data in all formats from all locations in both physical and digital formats, as relevant, once the purpose is achieved or the project is abandoned—whichever is earlier in time. University shall, and shall ensure Principal Investigator and its research team also shall, delete the Data within twenty-four (24) hours from the end of Term.

3.2 **No Further Transfer.** University will not transfer Data to any third party without the prior written consent from Company. The parties agree that no third-party processing shall be done without first executing a Data Processing Agreement.

3.3 **Reporting.** In consideration of Company having provided Data, University will report the results of the Research Program to Company.

3.4 **Compliance with Law.** University's use of Data will comply all applicable federal, state and local laws and regulations.

CONFIDENTIAL INFORMATION

4.1 **Definition of Confidential Information.** "Confidential Information" means confidential, scientific, business or financial information that is provided in written form and clearly marked as Confidential provided that such information:

(A) is not publicly known or available from other sources who are not under a confidentiality obligation to the source of the information;

(B) is not already known by or available to the receiving party without a confidentiality obligation; or

(C) is not independently developed by the receiving party.

4.2 **No Disclosure.** The receiving party will protect the disclosed Confidential Information by using the same degree of care, but no less than a reasonable degree of care, to prevent unauthorized use or disclosure of the Confidential Information as the receiving party uses to protect its own confidential information of a like nature.

4.3 **Confidentiality Term.** The receiving party's obligations of confidentiality will continue for five (5) years from the date of termination or expiration of this Agreement.

4.4 **Compelled Disclosure.** If the receiving party is required to divulge Confidential Information either by a court of law or in order to comply with any federal, state or local law or regulation, the receiving party will provide the disclosing party with reasonable notice.

PUBLICATION

Principal Investigator will be free to publish and otherwise publicly disclose the Results provided that the conditions in this section and in this Agreement are fulfilled. Principal Investigator shall provide to Company a confidential copy of any such proposed publication or disclosure at least thirty (30) days prior to publication. Within that thirty (30) day review period, Company may require Principal Investigator to delete any Company Confidential Information from the proposed publication, and, if the proposed publication contains patentable subject matter directly relating to the Data, then at Company's written request within said thirty (30) day period, the Principal Investigator will delay publication for up to an additional thirty (30) days to allow for filing of patent application(s). If negative findings are to be included in Principal Investigator's final report, thirty (30) days before submission of any proposed publication or presentation or at least five (5) days before submission of any proposed abstracts, Company reserves the right to take immediate action to begin remediation efforts and have any efforts to comply with recommendations be documented in such publications, presentations and/or reports of the Results. Notwithstanding anything to the contrary herein, Company agrees to allow Principal Investigator to publish and disclose sufficient information regarding the Data to enable the complete and accurate publication of Principal Investigator's Results. University and Principal Investigator will maintain all such prepublication materials in confidence in accordance with Section 4 ("Confidential Information") of this Agreement. The Principal Investigator will furnish Company with periodic written reports on the progress of the Research Program as mutually agreed by the parties and reasonably consistent with applicable research standards. The Results shall be formatted to meet the needs of diverse audiences, including AI researchers, I/O psychologists, HR analysts, Company's current and prospective clients, and stakeholders. Company may provide suggestions on the final format for the Results. The final draft of the Results shall include the purpose, methods and findings of the Research Program.

PUBLICITY

Neither party will use the name or trademark of the other party, or the names of the other party's employees, students or agents in any publicity, advertising or announcement related to this Agreement without the prior written consent of the other party's authorized officials.

GENERAL PROVISIONS

7.1 **No Warranties.** Except as stated in Section 2.2, Data are provided by Company AS IS, WITHOUT ANY WARRANTIES, EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION ANY WARRANTY OF FITNESS FOR A PARTICULAR PURPOSE.

7.2 **Liability.** In no event shall Company be liable for any use by University of Data or Results or for any loss, claim, damage, or liability, of any kind or nature, that may arise from or in connection with this Agreement or University's use, handling, or storage of Data. University agrees to indemnify and hold harmless Company, its trustees, officers, employees, students, volunteers and agents from all liability, loss, or damage they may suffer as a result of claims, demands, costs or judgments against Company arising out of the use, handling or storage of Data by University. University will ensure that Principal Investigator complies with all aspects of this Agreement and shall be responsible for any breach of this Agreement by Principal Investigator.

7.3 **Termination.** Either party may terminate this Agreement at any time upon thirty (30) days prior written notice, in which case University will discontinue within thirty (30) days use of the Data and related information. University agrees, upon Company's direction, to return or destroy Data. Sections 2.1, 3.1, 3.2, 3.4, 4, 5, 7.1, 7.2 will survive the termination or expiration of this Agreement.

7.4 **Notice.** All notices under this Agreement are deemed fully given when written, addressed, and sent as follows:

All notices to Company are e-mailed or mailed to:
pymetrics, Inc.
Legal Department
Address

All notices to University are e-mailed or mailed to:
[University Contracts Office]
Address

cc: Principal Investigator

7.5 **Severability.** If any paragraph, term, condition or provision of this Agreement is found by a court of competent jurisdiction to be invalid or unenforceable, or if any paragraph, term, condition or provision is found to violate or contravene the substantive laws of the State of [State], then the paragraph, term, condition or provision so found will be deemed severed from this Agreement, but all other paragraphs, terms, conditions and provisions will remain in full force and effect.

7.6 **Integration.** This Agreement, including attached Exhibits, supersedes all prior oral and written proposals and communications, if any, and sets forth the entire agreement of the parties with respect to the subject matter hereof, and may not be altered or amended except in writing and signed by an authorized representative of each party.

7.7 **Electronic Copy.** The parties to this document agree that a copy of the original signature (including an electronic copy) may be used for any and all purposes for which the original signature may have been used. The parties further waive any right to challenge the admissibility or authenticity of this document in a court of law based solely on the absence of an original signature.

The duly authorized party representatives execute this Agreement.

[University]	**COMPANY**
Signature:	Signature:
Name: [Redacted]	Name: [Redacted]
Title: [Redacted]	Title: [Redacted]
Date: [Redacted]	Date: [Redacted]

I acknowledge that I have read this Agreement in its entirety and will use reasonable efforts to uphold my obligations and responsibilities under this Agreement.

PRINCIPAL INVESTIGATOR
Signature:
Name: [Redacted]
Title: [Redacted]
Date: [Redacted]

Exhibit A – Research Program

Algorithmic hiring is a broadly deployed form of algorithmic decision-making. To implement algorithmic hiring, many firms rely on vendors of hiring algorithms, such as pymetrics. Given the stakes of hiring/employment, and the possible risks of algorithmic decision-making, naturally there are questions of fairness and equity: how do these systems perform across different protected categories and demographic subgroups?

In our study, we aim to further analyze the nature of the pymetrics' algorithms to see how they perform across different deployments. Concretely, our work will center on *outcome homogenization*, a type of systemic harm. That is, we will analyze whether the pymetrics algorithms have a tendency to rate the same candidates highly across all systems they would provide to clients and the same candidates lowly across all systems. We will also perform various group-level analyses in addition to this, and more generally understand a variety of individual-centric and group-centric phenomena.

Table 3. **Descriptive Statistics for Applications Based on Self-Reported Employer Metadata**.

Category	%
Industries	
Professional Services	22.37
Financial Services	17.11
Manufacturing	10.53
Technology	9.21
Other Industries	40.79
SOC Codes	
Hand Laborers and Material Movers (53-7062)	12.67
Financial and Investment Analysts (13-2051)	5.78
Customer Service Representatives (43-4051)	4.12
Software Developers (15-1252)	2.46
Missing	50.44
Account Region	
North America	55.26
Europe, the Middle East and Africa	25.00
Asia-Pacific	17.11
Missing	2.63
Client Revenue	
Over $5 bil.	51.32
$1 bil.–$5 bil.	14.47
$250 mil.–$500 mil.	3.94
$500 mil.–$1 bil.	2.63
Missing	25.00

A.2 Descriptive statistics

In §2, we report descriptive statistics for applicant metadata (see Table 1) along with information about pymetrics clients and the clients' positions with more information provided in Table 3. In addition, we report the counts for how many applications are submitted by each distinct application as it is central to our analysis of homogeneous outcomes and systemic rejection (see Table 4).

B Adverse Impact Analysis

We report additional empirical results on adverse impact. In the main paper, we present the results of our adverse impact analysis when disaggregating on a per-position basis and stratifying by race and O*NET code. In addition to those results, which we foreground because of the significant disparities for race, here we present results for gender (Figure 5) and for gender-race intersections (Figure 6). The gender results demonstrate that very few positions meet both criteria of interest (i.e. $r_g < 0.8; z_g > 1.96$) for both the Male and Female groups. The intersectional results demonstrate similar disparities to the underlying racial groups irrespective of the gender. More detailed statistics for all demographics groups are reported in Table 5.

Table 4. **Distribution of Number of Applications Across pymetrics-Mediated Positions**.

Number of Applications	Count	Value (%)
1	2,837,952	84.16%
2	368,192	10.92%
3	101,518	3.01%
4	35,712	1.06%
5	14,851	0.44%
6	6,631	0.20%
7	3,142	0.09%
8	1,680	0.05%
9	1,047	0.03%
10	522	0.02%
11	321	0.01%
12	201	0.01%
13	133	0.00%
14	88	0.00%
15	53	0.00%
16	30	0.00%
17	23	0.00%
18	11	0.00%
19	6	0.00%
20+	19	0.00%

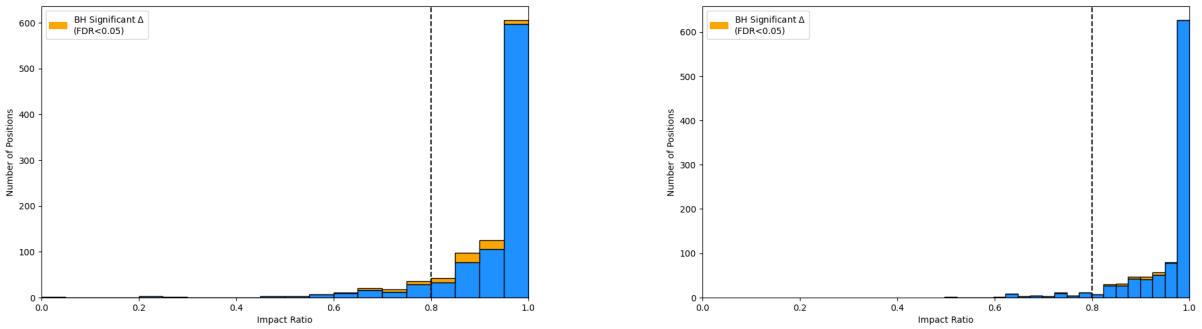
Table 5. **Adverse impact analysis by demographic group.** We report the aggregate selection rate and impact ratio for each demographic group (using the microaverage across positions) as well as the number and share of models that demonstrate adverse impact against that group (Panel A). Given the models that demonstrate adverse impact against a particular demographic group, we report the number and percentage of applications submitted to these models, the number and percentage of applicants that apply to at least one such model, and the shortfall and shortfall percentage (Panel B). We use "biased" to abbreviate adverse impact.

Panel A: Presence of Adverse Impact

Group	Aggregate Selection Rate	Aggregate Impact Ratio	Biased Models	Biased Models (%)
Asian	0.533	0.870	47	5.32
Black	0.525	0.839	82	10.62
Hispanic/Latino	0.568	0.916	6	0.80
White	0.583	0.962	4	0.46
Female	0.551	0.963	17	1.75
Male	0.557	0.982	1	0.10
Female Asian	0.519	0.808	57	6.70
Female Black	0.530	0.803	56	8.48
Female Hispanic/Latino	0.567	0.857	8	1.21
Female White	0.582	0.908	3	0.36
Male Asian	0.541	0.833	45	5.23
Male Black	0.520	0.808	71	9.67
Male Hispanic/Latino	0.568	0.874	9	1.29
Male White	0.583	0.922	4	0.47

Panel B: Effects of Adverse Impact

Group	Applications to Biased Models	Applications to Biased Models (%)	Applicants to Biased Models	Applicants to Biased Models (%)	Shortfall	Shortfall (%)
Asian	115317	14.74	105118	18.53	29320	3.75
Black	39986	25.87	36921	30.70	11513	7.45
Hispanic/Latino	2081	1.56	2063	2.03	647	0.48
White	5232	0.82	5232	1.09	1519	0.24
Female	4825	0.56	4736	0.72	1218	0.14
Male	194	0.02	194	0.02	49	0.00
Female Asian	59034	19.61	54723	24.88	16602	5.52
Female Black	29772	43.84	26805	50.78	7838	11.54
Female Hispanic/Latino	1436	2.72	1428	3.46	427	0.81
Female White	1609	0.67	1609	0.87	406	0.17
Male Asian	92502	19.47	84419	24.61	24434	5.14
Male Black	30158	35.43	26975	40.65	8610	10.12
Male Hispanic/Latino	3421	4.30	3275	5.50	978	1.23
Male White	3527	0.90	3527	1.21	1144	0.29



(a) Female (b) Male

Fig. 5. **Impact ratios per position by gender.** The histograms plot the impact ratio by gender group for each of the 1,746 disaggregated positions. We emphasize whether the impact ratio falls below 0.8 and whether the group is selected at a rate significantly lower than that of the most select group based on a two-sample pooled-proportion z-test for $p < 0.05$ subject to Benjamini-Hochberg correction. If a position does not receive any applicants from members of a particular demographic group, we do not visualize the impact ratio for that position for that group.

C Homogeneity Analysis

We report additional information on the methods we use to study homogeneous outcomes as well as additional results.

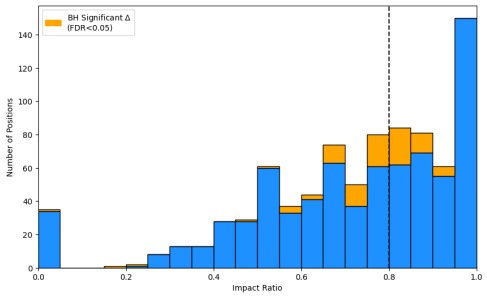
C.1 Methods

In the main paper, we provide the notation for computing the observed and baseline outcomes in the setting where N applicants each apply to same fixed set of k positions. Here we generalize this notation to consider N applicants that each apply k positions among M total positions. This framework is more general because applicants apply to the same fixed number of positions, but they may apply to entirely different positions. Each applicant will still receive some number of recommends from $0 - k$ and, as a result, the observed distribution of outcomes is computed in the same way. However, because the positions vary across applicants, the relevant baseline distribution will vary across applicants to reflect the specific positions they applied to. This generalization extends the prior work [57] that did not need to consider this complexity.

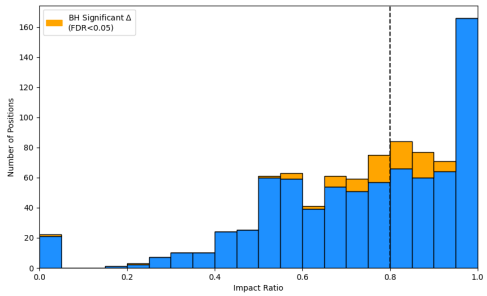
Define the application matrix $A \in \{0,1\}^{N \times M}$ such that $A[i,j]$ indicates if applicant i applied to position j where 1 indicates they applied and 0 indicates they did not apply. By definition, $\sum_{j=1}^{M} A[i,j] = k$ for all i. Define the outcome matrix $O \in \{0,1\}^{N \times M}$ such that $O[i,j]$ is the outcome for applicant i applying to position j, where 1 indicates recommendation and 0 indicates rejection. O is not defined where the given applicant did not apply to the given position. For $t \in \{0, \ldots, k\}$, the observed rate at which applicants receive t recommendations, and the baseline rate for t recommendations, are defined as follows:

$$P_{\text{observed}}(t \text{ rec.}) = \frac{\sum_{i=1}^{N} \mathbb{I}\left[t = \sum_{j=1}^{M} A[i,j]O[i,j]\right]}{N} \tag{3}$$

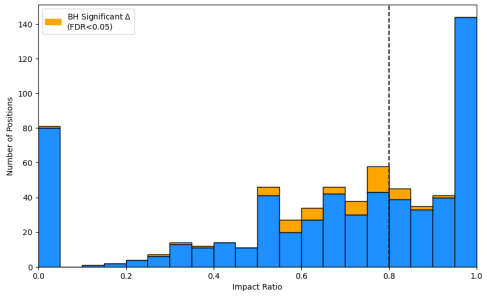
$$P_{\text{baseline}}(t \text{ rec.}) = \frac{\sum_{i=1}^{N} \text{Poisson-Binomial}(\{s_j | A[i,j] = 1\})[t]}{N} \tag{4}$$



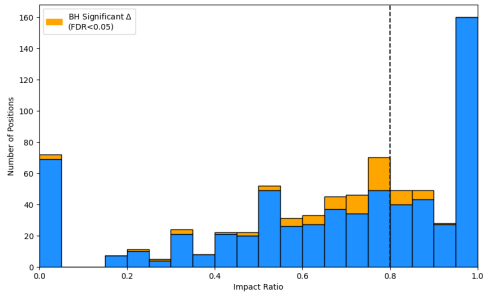
(a) Female Asian



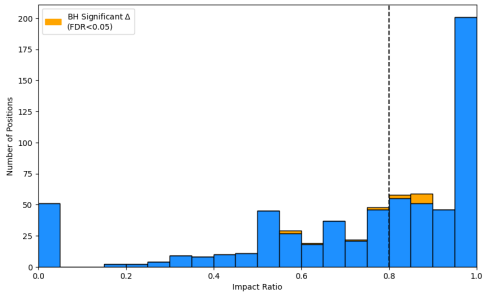
(b) Male Asian



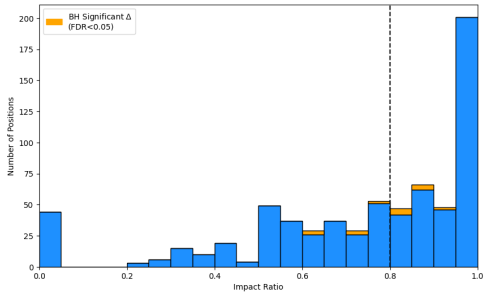
(c) Female Black



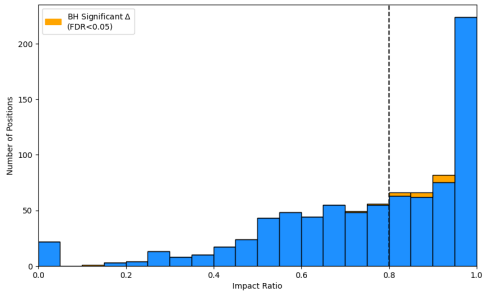
(d) Male Black



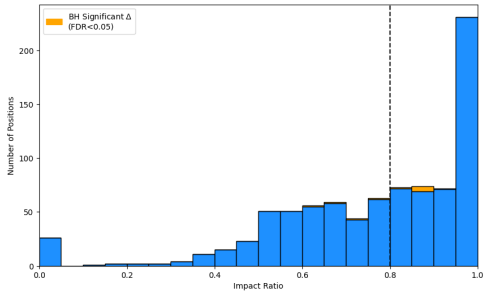
(e) Female Hispanic/Latino



(f) Male Hispanic/Latino



(g) Female White



(h) Male White

Fig. 6. **Impact ratios per position by (gender, race) intersection.** The histograms plot the impact ratio by (gender, race) intersectional group for each of the 1,746 disaggregated positions. We emphasize whether the impact ratio falls below 0.8 and whether the group is selected at a rate significantly lower than that of the most select group based on a two-sample pooled-proportion z-test for $p < 0.05$ subject to Benjamini-Hochberg correction. If a position does not receive any applicants from members of a particular demographic group, we do not visualize the impact ratio for that position for that group.

Table 6. **Homogeneous outcomes in Kline et al. (2022) data.** The observed and baseline systemic rejection rates in the Kline et al. correspondence study [38] involving 108 US companies.

Number of applications	Count	Baseline	Observed
1	8,209	76.0%	75.5%
2	3,105	57.7%	58.2%
3	1,070	43.8%	41.6%
4	318	33.3%	28.9%
5	65	25.3%	27.7%
6	21	19.2%	14.3%
7	21	14.6%	4.8%
8	39	11.1%	12.8%
9	71	8.4%	5.6%
10	124	6.4%	3.2%
11	193	4.9%	6.2%
12	279	3.7%	4.7%
13	346	2.8%	3.8%
14	355	2.1%	1.1%
15	375	1.6%	1.9%
16	405	1.2%	1.5%
17	417	0.9%	0.5%
18	340	0.7%	0.6%
19	281	0.5%	0.7%
20	218	0.4%	0.5%
21	176	0.3%	0.0%
22	132	0.2%	0.8%
23	91	0.2%	0.0%
24	72	0.1%	0.0%
25	25	0.1%	0.0%

C.2 Figures on Homogenization

The following figures demonstrate the observed and baseline number of model rejections based on the pymetrics data.

The following figure shows the homogeneous outcomes for the simulation exercise in §3.4.

C.3 Additional results from Kline et al.

Beyond studying pymetrics data, we also study homogeneous outcomes in data from Kline et al. [38]. In Table 6 we report the underlying observed and baseline systemic rejection rates that we visualize in Figure 3.

Received 13 January 2026

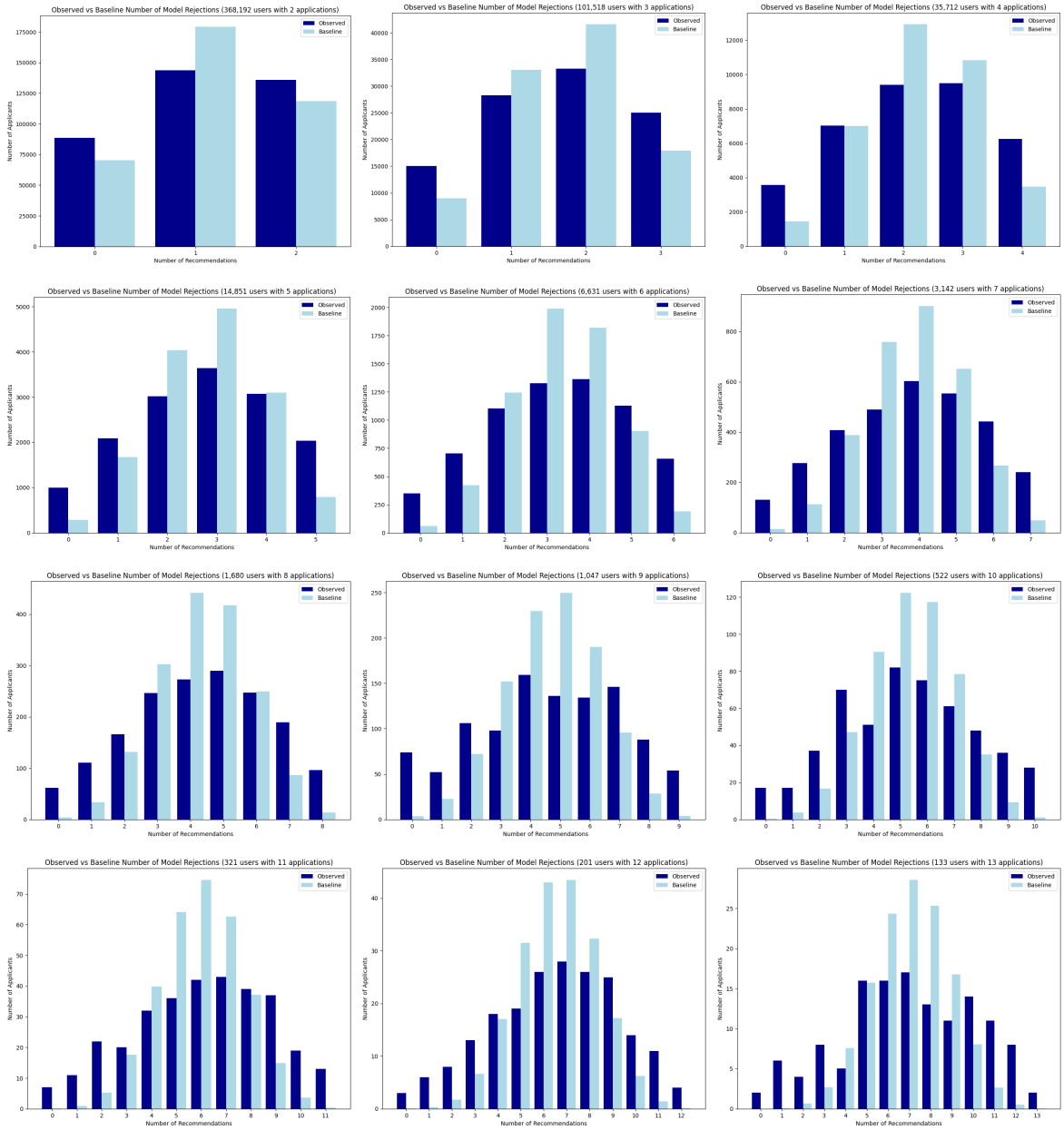


Figure 7. **Homogeneous outcomes by number of applications.** Subfigures depict the outcomes for applicants that submit exactly 2 through 13 applications: applicants are considerably more likely to receive homogeneous outcomes than under the baseline.

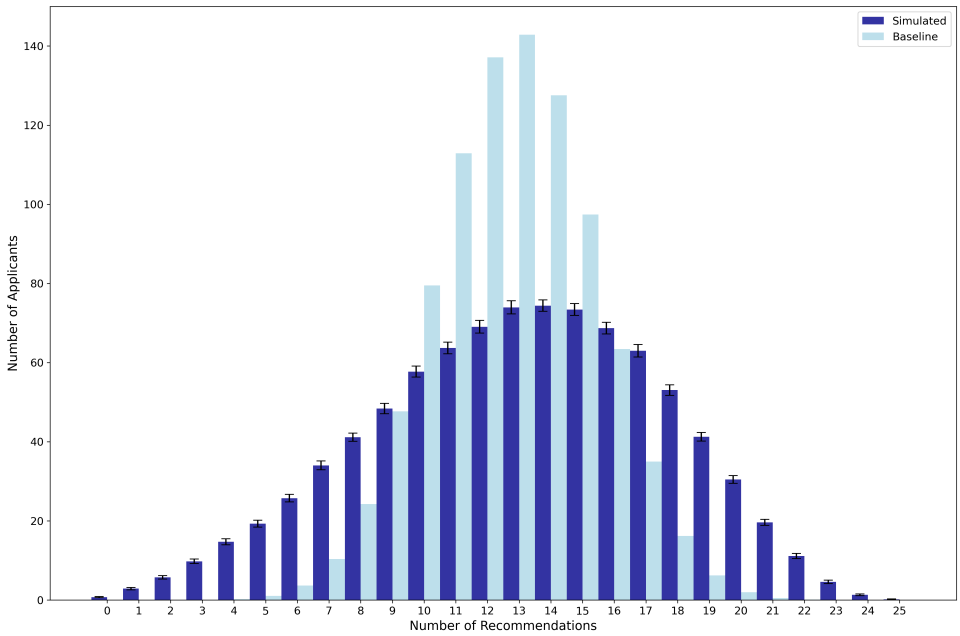


Fig. 8. **Observed vs Baseline Number of Model Rejections in the Connected Set** Given the observed set S of models an applicant is assessed by in reality, the *connected set* $S' \supseteq S$ contains every model that assessed an applicant also assessed by a model in S. Applicant outcomes for 25 randomly sampled models in S'. Confidence intervals are based on 100 simulations.