Information Processing and Management 63 (2026) 104506

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Does fair ranking lead to fair recruitment outcomes? A study of interventions, interfaces, and interactions

Alessandro Fabris [a,b,*], Clara Rus [c], Jorge Saldivar [d], Anna Gatzioura [d], Asia J. Biega [a], Carlos Castillo [d,e]

[a] Max Planck Institute for Security and Privacy, Bochum, Germany
[b] University of Trieste, Trieste, Italy
[c] Universiteit van Amsterdam, Amsterdam, Netherlands
[d] Universitat Pompeu Fabra, Barcelona, Spain
[e] ICREA, Barcelona, Spain

ARTICLE INFO

Keywords:
Algorithmic recruitment
Information access systems
Fairness in ranking
Fairness in outcomes
Position bias

ABSTRACT

Personnel recruitment is increasingly mediated by Applicant Tracking Systems (ATS), which rank candidates for job positions, making them a central decision-support tool in modern Human Resources (HR) processes. Often framed as an information retrieval (IR) problem, the ranking of candidates in ATS is typically driven by relevance to the job position, with algorithms sorting applicants according to a set of predefined criteria. In recent years, fairness-aware ranking methods have emerged to mitigate the risk of indirect discrimination, where the ordering of candidates may inadvertently favor one demographic group over another. These approaches are inspired by browsing models developed for web search and aim to balance candidate exposure based on protected characteristics. However, ATS in recruitment introduce unique challenges due to their high-stakes nature and the decision-making context in which they operate. In this paper, we present a series of user studies that explore the disconnect between *fair exposure* and *fair outcomes* in candidate shortlisting. We focus on how factors such as task design (e.g., how recruiters interact with candidate lists), individual representations of candidates (e.g., national origin cues), and ranking order influence both position bias and demographic balance. Our findings show that while demographic balance may be achieved in terms of ranking visibility, this does not necessarily translate to fair outcomes in terms of who gets shortlisted. Through a crowdsourced experiment and in-depth interviews with recruiters, we identify key task-level, individual, and ranking factors that mediate these effects. We conclude that fairness in ATS rankings is contingent not only on algorithmic design but also on the shortlisting tasks they support, as well as the interfaces, strategies, and assumptions that recruiters use when interacting with candidate lists. Based on these insights, we provide implications for the design of algorithms, interfaces, and recruitment processes that support fairer and more equitable recruitment outcomes.

1. Introduction

Corporate job postings routinely attract hundreds of applications, often exceeding 200 per opening (Fuller et al., 2021), making Applicant Tracking Systems (ATS) an essential tool for managing the recruitment pipeline. These systems structure the recruit-

* Corresponding author.
E-mail addresses: alessandro.fabris@units.it (A. Fabris), c.a.rus@uva.nl (C. Rus), jorge.saldivar@upf.edu (J. Saldivar), anna.gkatzioura@upf.edu (A. Gatzioura), asia.biega@mpi-sp.org (A.J. Biega), chato@icrea.cat (C. Castillo).

https://doi.org/10.1016/j.ipm.2025.104506
Received 9 May 2025; Received in revised form 20 November 2025; Accepted 21 November 2025
Available online 8 December 2025

ment workflow, enabling HR teams to efficiently process large volumes of applicants through stages such as longlisting (identifying a broad pool of suitable candidates) and shortlisting (selecting a small set for interviews or more intensive evaluation) (Huang et al., 2023).

To make recruitment manageable, key functionalities of an ATS include selecting and ranking candidates to create lists that recruiters examine. Despite claims of reduced bias (Kappen & Naber, 2021; Miller, 2018), selection and ranking algorithms run the risk of systematically downgrading vulnerable candidates when deployed, leading to undesirable discrimination (Dastin, 2018; Geyik et al., 2018). Candidate ranking is usually framed as an information retrieval (IR) problem, where candidates are sorted according to the probability ranking principle (Robertson, 1977), placing the candidates who are more likely to be relevant at the top. However, the analogy to document retrieval has important limitations. ATS are not merely retrieval tools–they are decision support systems that assist humans in making consequential judgments about people. Unlike documents, candidates have rights and interests that raise legal, ethical, and social concerns (Rigotti & Fosch-Villaronga, 2024). These concerns are not just theoretical. In Europe, for instance, systems "intended to be used for the recruitment or selection of natural persons" and "to analyze and filter job applications" are considered *high-risk systems* and hence subject to increased requirements and oversight when powered by artificial intelligence (European Union, 2024).

In response, recent research has focused on *fairness-aware ranking algorithms*, which aim to mitigate disparities in visibility—and, by extension, selection probabilities—between candidates from different demographic groups (Amigó et al., 2023; Bigdeli et al., 2022; Zehlike et al., 2022a,b). This line of work draws from insights in web search (Craswell et al., 2008), where *position bias* (also known as "ordering effect") has long been recognized: users tend to focus disproportionately on top-ranked items. In the recruitment context, if a ranking algorithm consistently places members of a protected group lower in the list, these candidates are systematically disadvantaged, regardless of qualifications. A core assumption underlying fairness-aware ranking algorithms is that allocating visibility equitably across demographic groups will lead to equitable selection rates. However, this assumption derives from browsing models developed for web search and holds only under idealized conditions. These models often overlook key contextual and cognitive factors that shape decision-making in recruitment settings.

Research Questions and Contributions. Motivated by this gap, we investigate the drivers of group-level selection outcomes in a controlled simulation where participants are tasked with shortlisting candidates for a job. We focus specifically on *demographic balance*, defined as the extent to which selection rates are statistically independent of candidate demographics. We examine three broad categories of factors that may influence both position bias and demographic balance. *Task-level factors* include characteristics such as the number of candidates to be selected, the mode of interaction with the ranking list (e.g., whether candidate details are revealed on click), and the difficulty of distinguishing among top-ranked candidates (e.g., how similar candidates are in terms of job fitness). *Individual representation factors* refer to attributes such as candidates' names, gender, experience, education, and skills–particularly whether names, prior employers, or educational institutions might be perceived as "foreign". *Ranking factors* refer to the ordering of candidates based on predicted job fit and the distribution of protected attributes across ranks.

To guide our investigation, we pose the following research questions:

- **RQ1.** How do task-level design factors shape the manifestation of position bias in candidate rankings?
- **RQ2.** How does position bias, in interaction with candidate representation and ranking characteristics, affect demographic balance in visibility and shortlisting?
- **RQ3.** How do recruiters perceive and navigate diversity when shortlisting candidates, and what strategies and assumptions influence their decisions?

To address RQ1 and RQ2, we conduct a large-scale experiment using a crowdsourcing platform, allowing us to systematically examine how task-level factors and ranking dynamics influence position bias and demographic balance. For RQ3, we complement this with a qualitative study involving think-aloud protocols and semi-structured interviews with 12 professional recruiters, providing deeper insight into the perceptions, assumptions, and strategies that underlie real-world shortlisting behavior. Overall, we make the following contributions:

- **Disconnect between fair exposure and fair outcomes**: we demonstrate that fairness-aware ranking, designed for equitable exposure across demographic groups, may not lead to fair shortlisting outcomes in recruitment contexts.
- **Empirical study of contributing factors**: through controlled experimentation and recruiter interviews, we uncover how task design, candidate representation, and recruiter behavior mediate the relationship between ranking position and candidate selection, shaping demographic disparities
- **Design implications for fair recruitment systems**: drawing on our findings, we outline concrete recommendations for algorithm, interface, and process design that promote more equitable recruitment practices.

Ethics Review. This study, including the research protocol, data protection aspects, and consent forms, was reviewed by the Ethics Review Board of Universitat Pompeu Fabra and after rounds of reviews and revisions, was approved.

Code and Data Release. The code and data used in our experiments, including the experimental UI, synthetic profiles, and job descriptions, are available at https://doi.org/10.5281/zenodo.17780217.

2. Related work

2.1. Discrimination in hiring

Much research has shown that biases, both explicit and implicit, shape recruitment outcomes. Gender and ethnicity, in particular, influence how candidates are evaluated, even when qualifications are kept constant (Bertrand & Mullainathan, 2004; Derous & Ryan, 2019; Koch et al., 2015; Peng et al., 2019; Simon et al., 2023; Steinpreis et al., 1999). Contextual factors, such as the recruiter's own demographics or the difficulty of the recruitment task, further complicate the picture (Peng et al., 2019). For example, when filling hard-to-staff positions, recruiters appear less sensitive to name-based signals of ethnicity (Baert et al., 2015), suggesting that the specific context of the labor market can moderate discriminatory behavior.

Experimental and audit studies have consistently demonstrated systemic discrimination based on race, gender, age, and other protected attributes, affecting outcomes such as interview callbacks, wages, and promotion rates (see Bertrand & Duflo, 2017; Neumark, 2018 for comprehensive surveys on the topic). In seminal work, Bertrand and Mullainathan (2004) showed that identical resumes with White-sounding names received 50 % more callbacks than those with African-American names. Such studies have been critical in establishing persistent patterns of inequality in employment decisions.

An explanation for asymmetric treatment in recruitment is rooted in *rational inattention*–the idea that decision-makers, constrained by cognitive limits, selectively attend to information using heuristics that may reflect underlying biases (Maćkowiak et al., 2023). The concept of *attention discrimination* models how recruiters may prioritize candidates from non-stigmatized groups under time pressure or ambiguity (Bartoš et al., 2016; Lahey & Oxley, 2018). In a nutshell, the propensity of decision-makers to acquire information on (and pay attention to) candidates from a group with negative stereotypes tends to decrease in competitive selection tasks. Although this has been documented in both lab and field settings, its implications in algorithm-mediated recruitment contexts–especially in ranked environments–remain underexplored.

2.2. Fairness in algorithmic hiring

Although AI-driven tools are often promoted as means to reduce bias and improve fairness in recruitment (Kelan, 2024), they can also encode or amplify existing inequalities (Raghavan et al., 2020). Bias may arise from training data, model assumptions, or deployment contexts (Ekstrand et al., 2022). Even without explicit access to protected attributes, algorithms can approximate them through proxies in resumes or social data, leading to unintended discrimination (De-Arteaga et al., 2019; Geyik et al., 2019).

Information access technology focused on ranking and recommendation is central in algorithmic hiring (Fabris et al., 2025). Technology assists recruiters by filtering promising candidates and presenting them in a suitable order (Chen et al., 2018; Geyik et al., 2019; Hannak et al., 2017; Wilson et al., 2021). Studies of professional search have shown that recruiters behave differently from general users. They explore deeper into ranked lists, conduct longer sessions, and rely on more complex queries (Kaya & Bogers, 2023). However, biases such as position and trust bias still influence their attention and selection decisions (Alvarez et al., 2025; Russell-Rose & Chamberlain, 2016), particularly under time pressure or information overload. Despite this, most research treats the human decision-maker as a passive agent or simplifies recruiter behavior using general-purpose browsing models drawn from web search (Li et al., 2020; Schumann et al., 2019). This risks overlooking how recruiters actually interact with ranked candidate lists.

The studies most closely related to ours focus on understanding whether promoting fairness in the ranking of the candidates impacts the recruitment outcomes of minority groups. Geyik et al. (2019) propose a bias mitigation technique and test it on the LinkedIn recruitment platform, demonstrating improvements in fairness metrics without affecting business metrics. The findings of Sühr et al. (2021) confirm that fair ranking algorithms can have a positive impact on the selection of minority groups.

Although prior work has provided important evidence of discriminatory recruitment outcomes, demonstrating that fair ranking holds the potential to mitigate them, several limitations remain. First, they focus on outcome disparities without modeling the underlying mechanisms. Second, they overlook important mediating factors, such as the assumptions and heuristics that recruiters use in the decision-making process. Finally, they lack actionable guidance to mitigate discrimination in recruitment systems. Our work addresses these limitations by modeling how attention and position bias influence discriminatory outcomes, conducting a qualitative analysis of recruiter behavior, and offering recommendations for the design of fairer algorithms, interfaces, and recruitment workflows. Additionally, our work focuses on ethnicity from a European perspective, supporting the operationalization of anti-discrimination principles in algorithmic hiring outlined in the EU AI Act (European Union, 2024).

3. Study design

Our study consists of a series of experiments in which participants select candidates for a job from a list. We vary task factors, representation factors, and the ordering in which candidates are shown. We build candidate profiles and job descriptions by carefully controlling for these variables. We measure the extent to which these factors affect disparate attention allocation through position bias and, ultimately, the demographic composition of the chosen candidates. In the description below, we highlight the experimental variables as **(EV)**.

3.1. Fairness notions

We study three notions of group fairness, focusing on ethnicity as a protected attribute in the context of recruitment in Europe. We model ethnicity according to European specificity (Section 3.3.1) and simplify it to a binary, distinguishing European candidates (EU – local and historically advantaged) from Non-European ones (NEU – foreign and disadvantaged). The decision space for fairness constructs in recruitment is, of course, very rich and nuanced (Fabris et al., 2025); here, we focus on these measures to compare prevalent exposure-based notions with (presumably related) outcome-based ones.

Exposure-based. First, in accordance with the literature on fair ranking (Biega et al., 2018; Singh & Joachims, 2018; Zehlike et al., 2022a), we measure the Relative Exposure Difference (`red`) for protected groups as

$$\texttt{red} = \frac{\text{Exp(EU)} - \text{Exp(NEU)}}{\text{Exp(EU)} + \text{Exp(NEU)}} \tag{1}$$

Here, $\text{Exp}(\cdot)$ is computed using a position-based browsing model with exponential decay (Moffat & Zobel, 2008), capturing the cumulative visibility of each group in ranked candidate lists. `red` ranges in $(-1, 1)$, with positive values indicating greater exposure for European candidates.

Outcome-based. Second, departing from typical fair ranking measures, we consider the distribution of *actual outcomes* (shortlisting decisions) that result from a ranking.[1] Relative Demographic Disparity (`rdd`) captures differences in shortlisting probability between demographic groups, inspired by the fairness criterion of independence (Barocas et al., 2023):

$$\texttt{rdd} = \frac{\Pr(\text{short}|\text{EU}) - \Pr(\text{short}|\text{NEU})}{\max\left(\Pr(\text{short}|\text{EU}), \Pr(\text{short}|\text{NEU})\right)}. \tag{2}$$

Here, positive values indicate disproportionate favoring of European candidates. The normalization accounts for differences in base rates, ensuring that large relative disparities are not masked by low absolute differences. Finally, we complement these constructs with a representational notion of fairness, *Difference in Representation* (`dr`), capturing diversity among the shortlisted candidates:

$$\texttt{dr} = \Pr(\text{EU}|\text{short}) - \Pr(\text{NEU}|\text{short}) \tag{3}$$

Given the equal distribution of demographics in candidate pools (Appendix A), `dr` aligns with both egalitarian and representational notions of diversity (Fazelpour & De-Arteaga, 2022).

3.2. Shortlisting tasks

Several task factors influence the competitiveness of a job opening, modulating recruiter attention mechanisms and position bias. We characterize competitiveness in terms of job type and selection rate.

3.2.1. Jobs

We select ten job positions for which online recruitment is highly relevant by sourcing a list of 10 in-demand jobs from LinkedIn (Lewis & Mohapatra, 2023), InfoJobs, and Randstad,[2] distinguishing between two job types based on their **pool fitness variance (EV)**.

Low-variance positions typically require few educational qualifications, skills, and prior experience. Minimum requirements tend to be met by most applicants; pools of CVs exhibit lower variability in fitness. These jobs tend to require less skilled labor and involve little or no management of other people. In this category, we include `administrative assistant`, `customer service representative`, `driver`, `factory operative`, and `warehouse specialist`.

High-variance positions have more stringent requirements and usually receive applications with a higher degree of fitness variability. They often require specialized training and/or management experience. In this category, we include `frontend developer`, `marketing manager`, `pediatric nurse`, `project manager`, and `senior store associate`.

We synthesize ten job descriptions, represented by requirements on (1) skills, (2) previous experience, and (3) education. We query job portals such as LinkedIn and InfoJobs to compile a list of skills, professional experience, and education frequently requested for each job as well as alternative names by which a given position might be known. The central panel of Fig. 1 contains the job description for `frontend developer`. We assign each position to one of five countries in Western Europe: France, Germany, Italy, Spain, and the Netherlands. Each country is assigned one high-variance and one low-variance job.

3.2.2. Selection rates

We vary the selection rates for the shortlisting task through two factors.

The **size of applicant pools (EV)** is the number of candidates shown, $n = 10$ or $n = 20$.

The **length of shortlists (EV)** is the number of candidates that the participant must select, $\ell = 1$ or $\ell = 3$.

[1] It is worth highlighting the distinction between immediate algorithmic outputs, such as fitness scores and rankings, and the decisions taken with the support of these outputs, such as recruitment decisions. In the algorithmic fairness literature and practice, it is common to measure the former (Fabris et al., 2025; Hireview, 2022; Raghavan et al., 2020), calling it *outcome-based fairness*. Arguably, this is a misnomer, as these measures focus on estimates produced by algorithms, abstracting away from the deployment conditions and the outcomes they support. In this work, *outcome-based* denotes measures that capture shortlisting decisions.

[2] The latter two, through private correspondence.

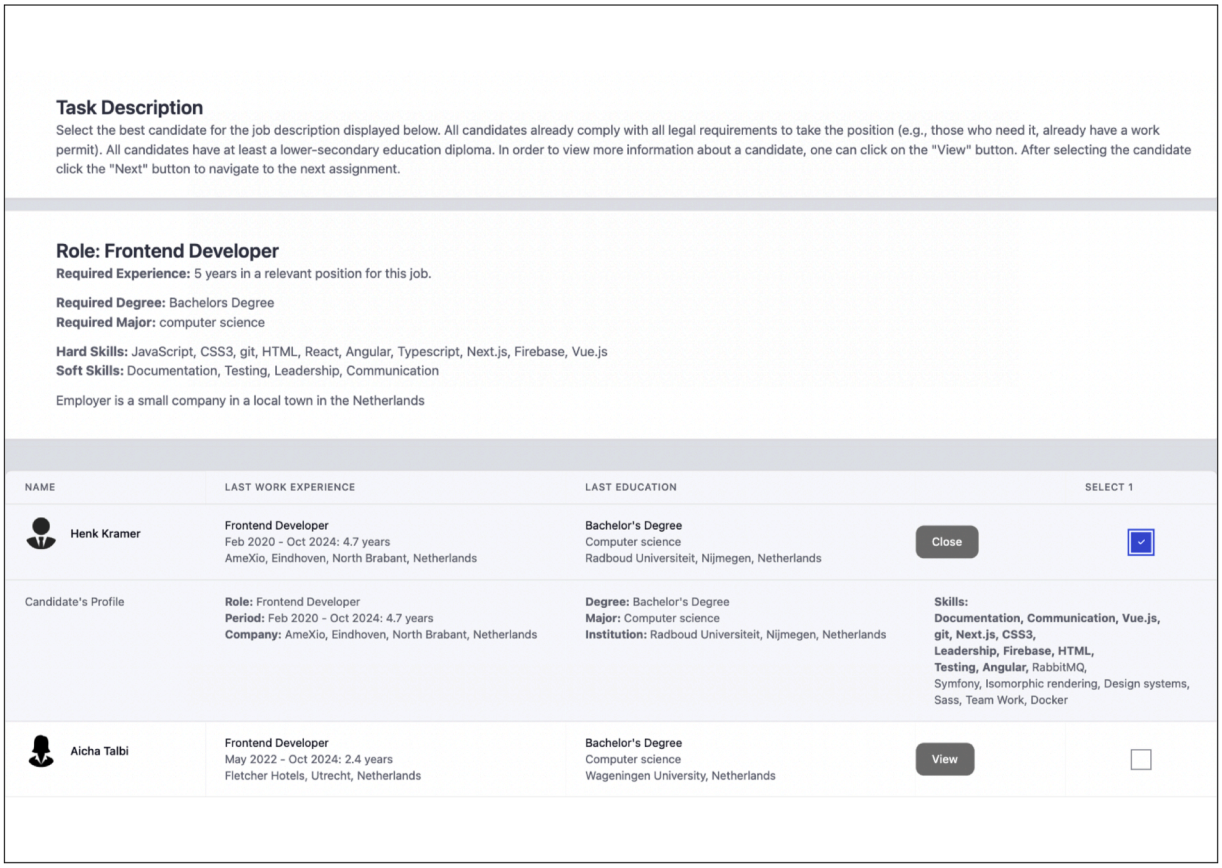

Fig. 1. User interface for the shortlisting task. Example of task description (top panel), job description (center), and candidate profiles (bottom).

3.3. *Candidate profiles*

We build 20 candidate profiles for each position. Profiles include past experience (employer, position, years), education (institution, degree), and skills drawn from real online job postings. This results in 200 candidate profiles, which are manually validated for consistency by at least one author. We manually associate real-world organizations with candidate profiles according to their experience, education, and *foreignness* (see Section 3.3.1). The bottom panel of Fig. 1 displays two profiles for `frontend developer`.

3.3.1. *Protected attributes*

Candidate profiles are associated with protected groups and convey this information through proxy features. Protected groups are defined by ethnicity (European, non-European) and binary gender (female, male).

Gender is expressed by candidates' forenames and by the presence of a gendered icon in the interface.

Ethnicity (EV) is conveyed by the forenames and surnames of the candidates. For candidates with non-European ancestry, we select names that are popular in the specified countries. For representativeness, we consider countries with sizable migration patterns to Europe (Eurostat, 2011, 2023); to convey non-European ancestry unambiguously, we focus on countries that have low linguistic overlap with European countries, especially with respect to names.[3] We source popular gendered forenames and surnames from a dedicated online service.[4] To avoid familiarity effects, we ensure that no surname repetitions occur within and between jobs.

Foreignness (EV). In addition to names, we use proxies for national origin in the work experience and education of candidates. Among the candidates of non-European ancestry, we simulate *long-term residency* and *recent immigration*. *Non-European long-term residents* are assumed to have lived in Europe their entire adult lives, and except for their names, their profiles are indistinguishable from European ones: their past employers and educational institutions are drawn from the same set.[5] *Non-European recent migrants*, on

[3] The non-European countries included in this experiment are Algeria, Bangladesh, China, Egypt, India, Ivory Coast, Morocco, Nigeria, Pakistan, Senegal, Somalia, and Turkey.

[4] https://forebears.io/

[5] We refer to these candidates as *non-European* to focus on ancestry and for simplicity of exposition. It is worth noting that they may qualify as European citizens under most national legal systems of the Union.

the other hand, have work experience and education in their country of origin. This is made explicit in their profile. Additionally, they use less common phrases to describe their experience and education, which are different from (but close synonyms of) those specified in job descriptions. This models the importance of keywords in recruitment (Team, 2025) and the fact that recent immigrants are more likely to present their experience in unusual terms, so their profiles are less likely to match the common keyword choices of recruiters (Fuller et al., 2021). Synonyms are sourced from online job postings; for example, the synonyms for "factory operative" include "production worker", "assembly operative", and "line operator".

3.3.2. Candidate fitness

We draw European and non-European candidates from the same fitness distribution by carefully designing their experience, education, and skills. We control for confounders, such as work gaps and age differences. We measure perceived candidate fitness by sourcing external ratings for each candidate. We hire crowdworkers through the Prolific online platform, with compensation of €10.5 per hour.

Each task involves reading a job description and scoring 10 applicants, in which names and organizations (employers and educational institutions) are redacted to mitigate potential biases. To encourage a thorough evaluation, we ask crowdworkers to provide separate fitness ratings for education, experience, and skills, based on candidate profiles matching job descriptions; finally, they provide an overall rating. All scores are on a 5-point Likert scale describing whether candidates: (1) do not satisfy any of the requirements, (2) satisfy some, (3) most, (4) all, or (5) exceed the requirements. The final candidate score is the average of ratings provided by 3 crowdworkers. Additional details are presented in Appendix A.

Overall, candidate profiles are a combination of education, work experience, and skills. To ensure that these characteristics represent desirable workers from the employer's perspective, we only include majors, job positions, and skills that appear as requirements in actual openings posted in online job portals (Section 3.2.1). This is a simplification that supports our research goals, as discussed in Section 5.3.

3.4. Ranking algorithms

We use fitness ratings to order candidates using three **ranking algorithms (EV)**. We leverage one fairness-unaware algorithm and two symmetrical algorithms that incorporate demographic constraints into the ranking process.

The **fitness-based** ordering (*fit*) ranks candidates by decreasing fitness for the job description.

The **discriminatory** (*discr*) and **positive action** (*pos*) orderings are built using the following procedure (Yang & Stoyanovich, 2017): split candidates into a European and non-European list, rank each list separately by decreasing fitness, and merge them from the top choosing with a different probability a European (p_{EU}) or non-European ($1 - p_{EU}$) candidate, until both lists are exhausted. The discriminatory ranking sets $p_{EU} = 0.8$, while the positive action ranking uses $p_{EU} = 0.2$. These simple ranking strategies exemplify the impact of systematically promoting or demoting candidates based on specific demographic attributes, an approach that lies at the core of more sophisticated fairness-aware ranking methodologies. We post-process the ranking to ensure that the top candidate is from the favored group (EU for *discr*, non-EU for *pos*) and that there is at least one non-favored candidate between the first and last candidate from the favored group.

3.5. Interface & interaction

The User Interface (UI) used for the experiment is an interactive web-based application that displays a task description, a job description, and a list of candidates (Fig. 1). The task description instructs participants to select $\ell = 1$ or $\ell = 3$ candidates and explains how to navigate the interface. The job description includes the role, the required experience, the degree, as well as hard and soft skills. The list of candidates displays the profiles with a gender icon, name, work experience, education, and skills.

We vary two factors that are known to influence human interaction with ranked lists and subsequent attention allocation (Agarwal et al., 2019; Goddard et al., 2012; Islam et al., 2019).

AI priming (EV) activates the prior expectations of algorithmic intelligence. Under the *aware* condition, recruiters read that the list contains "the top *n* candidates selected by an AI out of 100 applicants." In the alternative condition (*unaware*), there is no information on how the candidates were selected/ordered; neither AI nor alternative mechanisms are mentioned.

Candidate details (EV) can be collapsed or expanded. When *expanded,* the UI directly displays full candidate profiles; recruiters can read all the relevant information (experience, education, skills) by simply scrolling down. When *collapsed,* the UI presents a summary for each candidate, including their name, their most recent work experience, and their last educational degree. Recruiters must click on a button to view previous work experience and other educational records.

3.6. Data collection for quantitative study

As is common in previous work (Sühr et al., 2021), we leverage a crowdsourcing platform for our quantitative study. We recruited 142 participants through the Prolific platform, who were compensated at an hourly rate of €10.50. All participants were based in European countries, English-speaking, and 18 years or older. This is a convenience sample which may not fully represent recruitment practices, as discussed under limitations in Section 5.3. We complement this quantitative experiment with interviews and think-aloud studies analyzing the strategies and assumptions of real-world recruiters (Section 3.7).

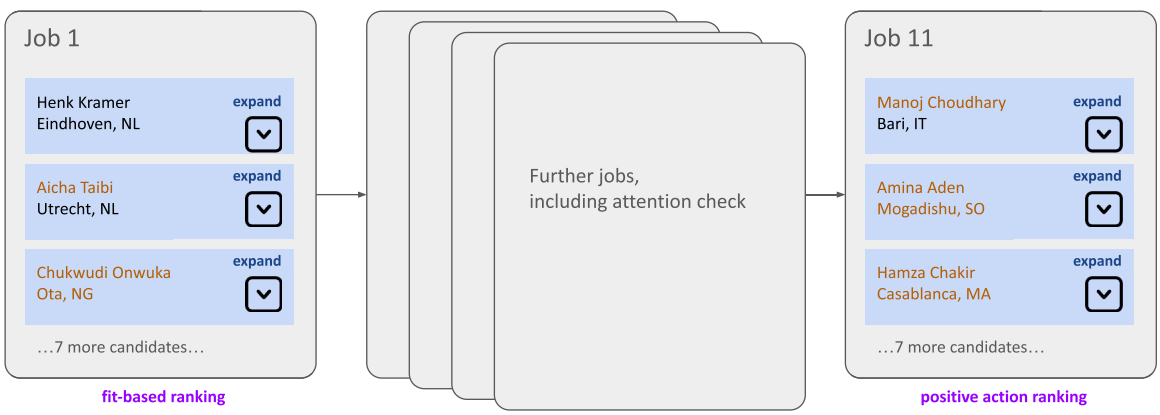

Fig. 2. Schematic overview of the experimental setup for a shortlisting session over multiple jobs. Fixed factors related to task design and interface configuration are shown in blue and remain constant throughout the shortlisting session. The ranking strategy, written in purple under the respective job, varies to mimic occasional fairness interventions. Elements conveying perceived foreignness, such as names and organizational affiliations, are highlighted in orange. In the UI, collapsed profiles contain basic information on work experience and education, which we omit from this schematic for simplicity. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2 summarizes the experimental variables. Each session involves shortlisting candidates for 10 jobs (plus 1 additional job serving as an attention check). Candidate pools have either high fitness variance or low fitness variance. Pools of $n \in \{10, 20\}$ candidates are shown, from which participants must shortlist $\ell \in \{1, 3\}$. Half of them are primed to consider the list as a ranking generated by an AI, while the remaining half are not primed to do so. Some see expanded candidate information, while others see it collapsed. Task and interface variables combine into 16 conditions; participants are assigned to one condition that remains fixed for them across all jobs. Fixed conditions are represented in blue in Fig. 2. The profiles represent European and non-European candidates; the latter can either have foreign-sounding names only or, additionally, experience/education at foreign organizations and use atypical keywords to describe their education and experience. Cues interpreted as markers of foreignness are colored in orange in Fig. 2. During a session, the ranking algorithms are rotated so that the same participant interacts with fitness-based, discriminatory, and positive-action rankings. We invite 3–4 participants for each combination of fixed and rotating conditions. If a participant fails the attention check, we discard the results of each task they performed. This leaves us with 2–4 valid results for each combination of job, ranking algorithm, and fixed condition. Overall, this corresponds to 10,000+ shortlisting outcomes. We pre-registered the quantitative experiment on position bias.[6]

3.7. Follow-up qualitative study

To enrich our quantitative findings, we conduct a qualitative study to uncover the perceptions, assumptions, and decision-making strategies employed by recruiters. This analysis aims to shed light on specific concerns that recruiters hold about foreign workers which may lead to inequitable treatment. This dual approach allows us to gain deeper insight into the factors that shape recruiter behavior and the underlying dynamics that contribute to imbalances in attention and outcomes recruitment. Furthermore, our analysis assesses the extent to which our experimental design mirrors real-world shortlisting practices and interface usage, ensuring both relevance and practical applicability.

Think-aloud and interview process. To complement our experimental results, we conducted a qualitative user study utilizing a think-aloud protocol and follow-up interviews with experienced recruiters. Participants completed a constrained version of the main task involving three job roles (`frontend developer`, `factory operative`, `project manager`), with candidate lists ($n = 10$) ranked by a fitness-based algorithm. Candidate profiles were collapsed by default and shown without AI disclosure. Recruiters selected the top candidate ($\ell = 1$), followed by exposure to two contrastive rankings–one discriminatory, one positive action–to elicit reasoning around potential downstream impacts of algorithmic bias. Sessions were 30 min long, conducted remotely, and are detailed in Appendix B These insights provide an interpretability context for model behavior and human decision-making in ranking interventions.

Participant profiles. Twelve participants (recruiters hereafter) completed the study. Most of the recruiters are female (8 out of 12), reflecting real-world demographics in HR management (Ainsworth & Pekarek, 2022), and their experience assessing job candidates ranges from a few recruitment processes (1 to 5) to more than 1000. Some recruiters have collaborated with an HR department and/or

[6] https://aspredicted.org/nwps-sbkz.pdf

Table 1
Demographics and recruiting experience of the participants in the qualitative study. All have participated in several recruitment processes, and additionally, some have collaborated with (Collab.) and/or worked within a Human Resources department (HR).

ID	Gender	Recruitment experience	Collab.	HR
P1	Female	51–200	✓	
P2	Female	1–5		
P3	Male	21–50		
P4	Female	6–20	✓	
P5	Male	6–20		
P6	Female	5–20		
P7	Female	1–5		
P8	Male	51–200	✓	
P9	Female	201–1000	✓	
P10	Male	51–200	✓	✓
P11	Female	21–50	✓	✓
P12	Female	1000 +	✓	✓

worked within one. Additional information on participants, including a summary of their demographics and recruitment experience, is reported in Table 1.

Interview analysis. The interviews were recorded and transcribed automatically using the video conferencing platform. Transcripts and researcher notes were consolidated into structured text documents for analysis. We used a deductive coding approach (Bingham & Witkowsky, 2021) to identify relevant themes related to recruiter behavior and perceptions. Details of the questionnaire and the coding procedure are provided in Appendix B.4.

4. Analysis & results

We first present the quantitative results on shortlisting fairness, followed by a qualitative analysis of recruiter assumptions on foreign and local candidates.

4.1. Quantitative study

4.1.1. RQ1: Positional bias

We investigate how task structure and interface design contribute to position bias in algorithmic rankings.

Measures. First, we define the *mean normalized rank* (`mnr`) for the shortlisted candidates as

$$\text{mnr}(\mathcal{E}) = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \frac{\bar{r}_i}{n_i - 1}, \tag{4}$$

where $\bar{r}_i$ is the average rank of the candidates shortlisted under condition i, n_i is the number of candidates, and $\mathcal{E}$ is the set of experimental conditions. This measure ranges from 0 to 1, with extreme values indicating that shortlisted candidates are selected from the top and bottom positions of the ranking, respectively. Second, we compute the average probability of shortlisting candidates in the bottom half of a ranking as follows:

$$\Pr_{\mathcal{E}}(\text{short}|\text{2nd half}) = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \Pr\left(r_i \geq \frac{n_i}{2}\right), \tag{5}$$

where r_i is the random variable for the rank of the candidates selected under experimental condition i. This definition is intuitive since a uniform shortlisting distribution would correspond to a 50 % probability, and smaller values indicate that candidates from the top ranks are selected more often.

Results. Table 2 summarizes the mean effects of each experimental factor on the mean normalized rank (`mnr`) and Pr(short|2nd half), with statistical significance assessed using paired-samples t-tests. Fig. 3 visualizes the impact on shortlisting rates for lower-ranked candidates.

Among the factors, pool fitness variance yields the strongest effect: candidates ranked in the bottom half are shortlisted 18 % of the time under high variance, compared to 31 % under low variance. Displaying candidate details also reduces position bias, increasing bottom-half selection rates by 5 percentage points when profiles are expanded. Position bias intensifies under lower selection rates—i.e., when fewer candidates are shortlisted from a larger pool. Although list length and shortlist size individually have smaller effects, they are statistically significant in `mnr` and result in a 3 % increase in Pr(short|2nd half).

Consistent with our pre-registered hypotheses, the likelihood of recruiters selecting candidates from the bottom half of the ranking remains below 50 % across all conditions but increases when the candidate pool has low fitness variance, when fewer candidates are shown, and when more candidates are to be shortlisted. These factors contribute to reducing recruiter reliance on rank. Contrary to expectations, expanded candidate details increase the likelihood of bottom-half selections. Although reaching lower-ranked profiles

Table 2
Influence of task and interface factors on position bias (RQ1).

	Mean norm. rank	Mean Pr(short| 2nd half)
Pool fit variance		
low variance	0.34	31 %
high variance	0.25	18 %
mean effect	+0.09	+13
p	< 0.0001	< 0.0001
Candidate details		
expanded	0.31	27 %
collapsed	0.29	22 %
mean effect	+0.03	+5
p	0.005	0.0007
N. of candidates		
10	0.32	26 %
20	0.28	23 %
mean effect	+0.04	+3
p	0.0008	0.06
Shortlist length		
3	0.31	26 %
1	0.28	23 %
mean effect	+0.03	+3
p	0.003	0.07
AI priming		
unaware	0.30	25 %
aware	0.29	24 %
mean effect	+0.01	+1
p	0.2	0.4

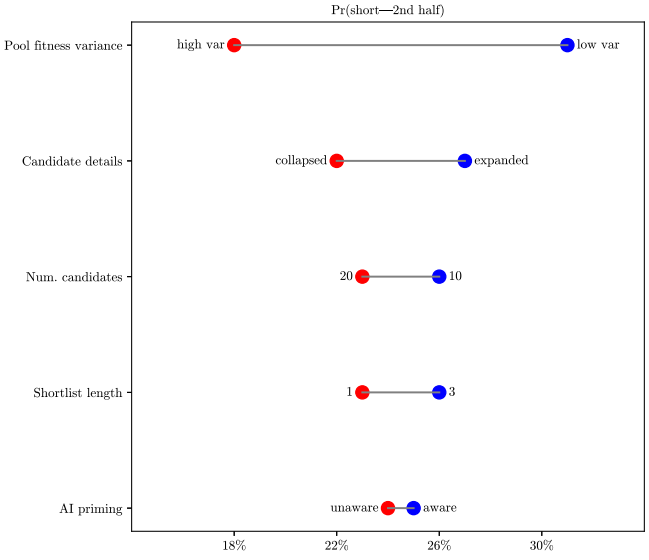

Fig. 3. Influence of task and interface factors on the probability of shortlisting items in the second half of a ranking (RQ1). Position bias is mitigated by low pool fitness variance, low UI friction, and higher selection rates (larger values, blue).

requires additional scrolling, removing the need to click to view full details appears to reduce interaction friction and promote broader exploration. We observe no significant effect of AI awareness (as operationalized in our priming condition) on position bias.

4.1.2. RQ2: Demographic balance

Next, we examine how position bias interacts with demographic attributes to affect shortlisting outcomes, considering both individual-level factors and ranking mechanisms.

Measures. To quantify the outcome imbalance, we use Relative Demographic Disparity (`rdd` - Eq. (2)) and Difference in Representation (`dr` – Eq. (3)). We further examine whether disparities in shortlisting outcomes correspond to disparities in visibility, by measuring the Relative Exposure Difference (`red` – Eq. (1)).

Results.

Table 3
Effect of the ranking algorithm on outcome imbalance (RQ2).

	fitness-based algorithm vs. balanced outcomes				discriminatory vs positive action algo			
	fit	balanced outcomes	mean effect	p	discr	pos	mean effect	p
rdd	0.12	0	+0.12	0.004	0.22	0.01	+0.21	<0.0001
dr	0.08	0	+0.08	0.02	0.17	0.00	+0.17	0.0002

Table 4
Effect of position bias and ranking algorithm on outcome imbalance (RQ2). Rankings are either discriminatory, or positive action. Low position bias is associated with balanced demographic outcomes.

	high position bias				low position bias			
	discr	pos	mean effect	p	discr	pos	mean effect	p
rdd	0.41	0.02	+0.39	<0.0001	0.04	0.00	+0.04	0.3
dr	0.34	0.00	+0.34	<0.0001	0.00	0.01	0.00	0.5

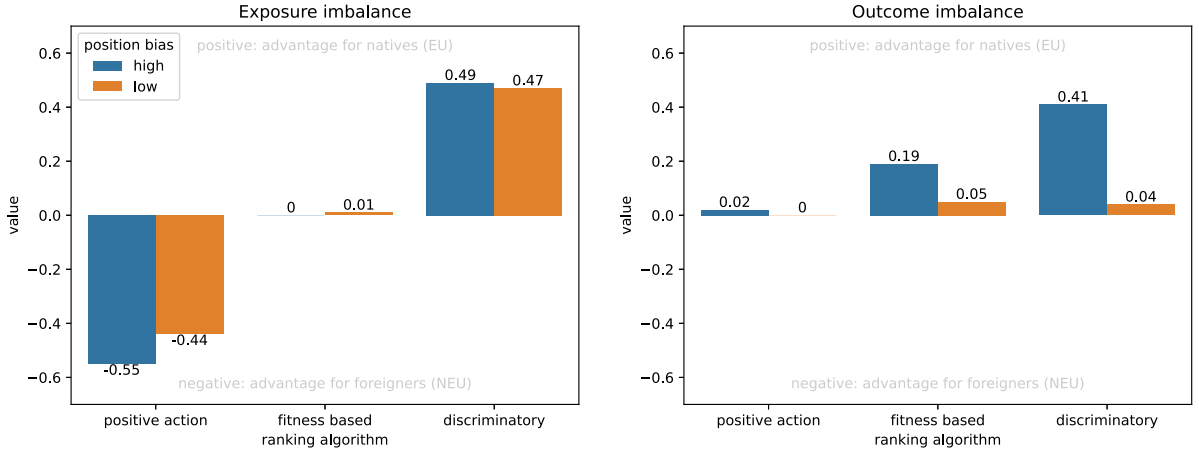

Fig. 4. Effect of position bias (color coded) and ranking algorithm (x axis) on exposure imbalance (red, left) and outcome imbalance (rdd, right) (RQ2). Low position bias is associated with balanced demographic outcomes (right). High position bias favours recruitment of local candidates (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3 summarizes the average outcome disparities under different ranking strategies. Fitness-based ranking leads to measurable imbalances favoring European candidates, with rdd = 0.12 and dr = 0.08, both statistically significant ($p < 0.05$, paired t-test). This indicates that European candidates not only have higher shortlisting probabilities but are also overrepresented among selected candidates. Discriminatory ranking amplifies these disparities, approximately doubling both metrics (rdd = 0.22, dr = 0.17). In contrast, positive-action ranking substantially reduces imbalance, yielding near-zero values (rdd = 0.01, dr = 0), though *without reversing the advantage*. The differences across ranking strategies are statistically significant, underscoring the role of algorithmic design in shaping fairness outcomes.

Next, we investigate the interaction between ranking algorithms and position bias by categorizing experiments into high and low position bias. Combinations of fixed experimental conditions and jobs that yield large values of mnr (Eq. (4)) across all ranking algorithms are labeled "high position bias"; conversely, experiments where mnr is below the median value are labeled "low position bias". Fig. 4 presents the results for rdd (right) and red (left) across ranking algorithms for both high and low position bias conditions As expected, red reflects strong exposure disparities under both discriminatory and positive action rankings, with near-zero exposure differences under fitness-based ranking. Notably, while demographic-based ranking strategies produce symmetric exposure advantages, *this symmetry does not necessarily extend to shortlisting outcomes*. For shortlisting outcomes (right panel), high position bias amplifies demographic imbalances, with discriminatory ranking yielding a significant advantage for European candidates (rdd = 0.41). Fitness-based ranking partially mitigates this effect (rdd = 0.19), while positive action ranking reduces the imbalance further but fails to advantage non-European candidates (rdd = 0.02). Table 4 reports the mean effects and statistical significance of the ranking algorithms on rdd and dr.

Finally, we investigate the impact of foreignness on the shortlisting of non-European candidates. Table 5 presents the comparison of mean shortlisting probabilities for candidates with low foreignness (i.e., long-term residents or second-generation migrants, identifiable only by their names) versus high foreignness (i.e., recent migrants, identifiable by their names, institutional affiliations, and keywords). For non-European candidates whose foreignness is inferred solely from their names, the shortlisting probabil-

Table 5
Differences between fitness scores and outcomes (probability of being shortlisted) between profiles representing non-EU candidates who are long-term residents versus recent migrants. The last column is for EU profiles, for comparison. Scores are on a 1–5 scale, indicating whether a candidate (1) does not satisfy, (2) satisfies some, (3) most, (4) all, or (5) exceeds the requirements. Differences in shortlisting probabilities are an average across all experiments.

	non-EU long-term residents	non-EU recent migrants	difference	%diff.	*p*-value	EU
average fitness	3.28	3.04	0.24	+7 %	0.02	3.26
Pr(shortlist)	0.18	0.12	0.05	+29 %	<0.0001	0.17

ity is Pr(short) = 0.18, which is similar to European candidates (Pr(short) = 0.17). However, when additional signals of foreignness are included (such as organization names and institutional affiliations), the shortlisting probability decreases by 29 %, dropping to Pr(short) = 0.12. This group of candidates also receives lower average ratings, as reported in the last row of Table 5, despite their fitness distributions being equivalent by design. Since candidate fitness was rated with redacted organizational details, the lower observed ratings are attributable to less common terms used to describe education and work experience.

4.2. Qualitative study

Finally, we study the human factors contributing to the observed patterns with a think-aloud study and semi-structured interviews; we focus on strategies, assumptions, and anticipated impacts of ranking strategies.

4.2.1. RQ3: Strategies, perceptions, and assumptions of recruiters

Selection Strategy. Recruiters in our study began by reviewing job descriptions to gain a deeper understanding of role requirements, such as experience, skills, and educational background, as well as the hiring organization. One participant explained this process as, *"So we're kind of just trying to get the context, with that information I am going to look at the individuals to see whether they meet the criteria or not"* (P1).

One participant mentioned that, when no "perfect match" could be found, they would look for the most "rounded" profile, while three participants paid special attention to the skill set of the candidates, *"The most deciding factor was their skills in general, their experience, even though I believe that someone can grow in experience"* (P6), "[…] *as long as they have the skills "* (P11), *"I would skills over experience"* (P9). Three of the participants mentioned the importance of education and would prioritize those with a higher level of education or those who have studied at prestigious institutions, *"the first thing I look at is the university people graduated"* (P5).

The majority of the participants followed a top-down browsing strategy when examining the candidate profiles (Granka et al., 2004), while two of them used a mixed top-down and bottom-up examination. In some cases, as soon as strong candidates are identified, they become a reference; recruiters pre-select them by using the checkboxes of the UI or by memorizing key aspects of their profile. Reference candidates are compared against the others and might be replaced if better candidates are found. One participant explicitly sought candidates who had studied at familiar or prestigious institutions to begin the examination. Six of them mentioned that company location and size were important factors *"size of the organization defines the working mindset"* (P12), while four would look for candidates already based at the company location *"local candidates* […] *would integrate better* […] *relocation can be considered as a last resort because it is usually risky and costly"* (P10).

Finally, four participants mentioned that neither the ranking of the candidates nor their demographic information would have an effect on their selection process as what they put emphasis on is to find the perfect candidate for a given position *"I'll have to look at all the profiles, just because somebody is at the top or at the bottom I cannot reject them or select them. I'll definitely have to go through the list and all the details again"* (P5). One of the recruiters mentioned that they would explicitly look for candidates from underrepresented groups (P11).

Assumptions. Recruiters brought a range of assumptions into the shortlisting process—about job location, hiring organizations, candidate attributes, and the ranking system itself—that subtly shaped their decisions and introduced bias, particularly against marginalized candidates.

Perceptions of cultural fit played a central role. Some recruiters assumed that local candidates would be more successful because of their familiarity with the language and norms. P7 stated, "So I also assume that the language would be important," and others wondered whether the candidates needed to understand location-specific regulations. Similarly, a participant speculated that candidates from large companies might struggle to adapt to small organizations, highlighting an assumption about organizational compatibility.

Assumptions about demographic identity were common. Recruiters often used names and educational backgrounds as cues for ethnicity, religion, or social origin. These perceptions influenced selection: "Mustafa Hussein […] could be discriminated because of the name" (P2), while P4 noted, "Wendy could have an advantage given the name." These comments reflect deeply embedded social biases and demonstrate how visible cues can affect perceived suitability.

aThree participants suggested that removing names might help mitigate such bias by de-emphasizing demographic signals and focusing attention on skills. However, assumptions did not only disadvantage candidates. In a rare instance of "positive" bias, one recruiter stated, "People from lower-income countries are more likely to work hard and at a lower salary; people from non-EU

backgrounds may be less demanding in relation to their rights" (P6), revealing how even seemingly favorable assumptions can be rooted in exploitative logics.

The participants also speculated on the mechanism behind the candidate ranking. Most assumed that the list was sorted by years of experience or the number of matching skills. In reality, it was based on a composite fitness score combining multiple dimensions (Section 3.3.2).

Impact of discriminatory and positive action rankings. At the end of their interview, participants were presented with two versions of the ranked candidate lists, generated by a discriminatory strategy and a positive action strategy, and were explicitly told about the imbalance in candidate distribution. Aware of this fact, participants were asked how each might affect recruiter behavior. Their reflections centered on how ranking order intersects with demographic visibility. Five participants underscored the strong influence of ranking position, pointing out that top-listed candidates are more likely to be closely reviewed due to recruiters' limited time and cognitive bandwidth: "The first person you see that fits the requirements […] would be okay, especially under time pressure" (P7). This aligns with the well-documented role of position bias in user interaction with ranked content.

Two recruiters suggested that positive action rankings could reduce bias, precisely because they prioritize candidates from underrepresented groups, making them more visible and therefore more likely to be considered. "If people are ordered using this [positive action] ranking, I think they [recruiters] would look more at their skills and choose based on their skills […] You'd have to go a lot to find someone from Europe […] and probably you will be scrolling the list until you find someone from Europe" (P6). These perspectives explain the quantitative results in Section 4.1.2 by highlighting the asymmetrical allocation of attention between groups.

5. Discussion

We summarize the main findings from our experiments, situating them in HR management, psychology, and algorithmic fairness scholarship. In summary, the quantitative results highlight a disadvantage for minority candidates, while the qualitative results describe some of the assumptions and cues that recruiters may leverage, which may contribute to this asymmetry.

5.1. Summary and interpretation of results

Disconnect between exposure and outcomes. Our findings reveal a gap between ranking position and recruitment outcomes, showing the complexity of extending information retrieval models to recruitment (Carterette, 2011; Craswell et al., 2008). Our results challenge the common assumption underlying fair ranking metrics, such as exposure disparity (Biega et al., 2018; Diaz et al., 2020; Singh & Joachims, 2018), that equalizing visibility improves outcomes (Section 4.1.2). Despite increased exposure, disadvantaged candidates are not shortlisted at equal rates, showing that exposure fairness does not always yield fair recruitment outcomes. This finding echoes (but differs from) the disconnect between different fairness notions, such as procedural and outcome-based fairness (Seppälä & Małecka, 2024). When rankings inform human decisions, which is most often the case, undesirable recruiter biases can influence outcomes (Almeida et al., 2019; Derous & Ryan, 2019) despite technical efforts of algorithmic balancing. This highlights the need for fairness-aware ranking algorithms to incorporate recruiter behavior, decision heuristics, and task constraints, motivating future work at the intersection of learning-to-rank and human-in-the-loop systems for responsible decision support in recruitment.

Perceptions and assumptions about foreign candidates. Our qualitative study highlights how recruiter assumptions about foreign candidates drive asymmetric outcomes (Section 4.2.1). Recruiters express concerns about perceived relocation costs, unfamiliarity with employers, relevance of education, language barriers, and a lack of understanding of relevant regulations. Some of these factors have been shown to penalize foreign candidates during recruitment due to a reduced perception of fit (Almeida et al., 2015). Names and affiliations often signal varying degrees of "foreignness," reinforcing these biases and affecting decisions (Section 4.1.2). Ethnic-sounding names and affiliations are important cues in the HR management literature as they are typically available with candidate profiles (Bertrand & Mullainathan, 2004; Derous et al., 2009), who may have to engage in concealment to mitigate systemic disadvantages (Kang et al., 2016). In this context, even ostensibly neutral fitness-based rankings can entrench inequality, as they reflect and reproduce past hiring decisions based on perceived (not actual) candidate fitness.

Asymmetry and position bias. The gap between fair rankings and fair outcomes is most evident under strong position bias. Even when non-European candidates are promoted to the top through positive action rankings, shortlisting advantages remain limited. In contrast, outcomes are more balanced when position bias is low (Section 4.1.2). This asymmetry can be explained through attention discrimination (Bartoš et al., 2016), where recruiters only engage with candidates who exceed an internal threshold of expected utility. Both foreignness and worse rank reduce a candidate's perceived utility and their likelihood of receiving attention. Position bias is associated with factors, such as lower selection rates, that make recruiters adopt a more strict comparison threshold. Non-European candidates often fall below this threshold. In high-bias scenarios–due to UI friction, cognitive load, or time pressure–recruiters narrow their attention to familiar and top-ranked candidates. In low-bias settings, on the other hand, the threshold drops. Factors like higher selection rates or fluid browsing encourage more thorough exploration. Recruiters are more likely to consider candidates with unfamiliar profiles and examine the bottom of rankings, leading to fairer outcomes.

Effects of AI awareness priming. The AI aware condition, in which participants of the quantitative study read that the list contained "the top n candidates selected by an AI out of 100 applicants" did not yield outcomes that were different from the AI unaware condition, in which they were simply shown the list. Previous work that examined perceptions of procedural fairness in HR processes described as making automated hiring decisions (Dineen et al., 2004; Lavanchy et al., 2023) or automated promotion/layoff decisions (Newman et al., 2020), consistently noted that human-made decisions were seen as more fair than automated decisions. Our qualitative study, in which several participants assumed that the list was automatically sorted according to some criterion, provides a possible explanation

of this apparent contradiction. We believe that human autonomy plays a role, and that the usage of an AI-sorted list from which a human selects candidates may be seen as substantially less problematic than having an AI select the candidates automatically.

5.2. Design implications

Modern ATS provide a central database for the recruitment efforts of a company. They store key information about job openings and possible candidates, helping recruiters assess candidate fitness and communicate with them (Leony et al., 2024). Given the high volume of applications for certain positions, candidate ranking and filtering are key functionalities of recruitment systems. Our findings shed light on the complexity of designing fair recruitment systems, emphasizing factors like perceived foreignness, cultural fit, and task design. Future work should focus on quantifying their impact on bias and decision-making. Given the challenge of exhaustively mapping all relevant factors, key immediate goals for designing fair systems include (i) encouraging deeper exploration by recruiters and (ii) postponing potentially bias-triggering interactions in the recruitment process. This section discusses design recommendations supporting these goals.

Designing interfaces. To promote exploration, shortlisting interfaces should reduce friction by displaying expanded profile details by default. This finding critically re-evaluates the assumed benefits of friction in system design, offering a counterexample of its effective scope and limitations (Rakova, 2023). ATS interfaces must strike a careful balance to avoid information overload and undesirable biases. Particularly, the presentation of demographic and institutional information (e.g., candidate name, gender, or educational background) should be delayed. This recommendation aligns with interventions to remove early barriers and, if not eliminate, delay the presentation of bias-inducing factors, supported by systems that improve job descriptions, pseudonymize candidate profiles, and design skill-based testing (Hsu, 2020; Vaishampayan et al., 2023; Yarger et al., 2020). Striking the right balance between showing essential information and reducing overload will help recruiters avoid fatigue and encourage more equitable decision-making.

Designing algorithms and evaluations. Algorithmic fairness initiatives in recruitment typically limit their focus to immediate algorithmic outputs, such as groupwise acceptance rates (Groves et al., 2024; Hireview, 2022) or ranking positions (Geyik et al., 2019). Our results show that equitable rankings and fair exposure do not always result in *actual* fair outcomes. Thus, it is crucial to monitor the recruitment process from start to finish. Fairness interventions should focus on outcome-based metrics rather than just ranking fairness. Additionally, task-specific models are needed that account for factors like recruiter browsing behavior (e.g., top-down vs. bottom-up) and position bias, moving beyond traditional models from web search.

Designing recruitment processes. Small changes in task design, such as increasing the number of candidates considered for shortlisting, can encourage recruiters to explore the candidate pool more thoroughly. Furthermore, support initiatives can assist job seekers, particularly non-native speakers or newcomers, in tailoring their CVs to relevant keywords and requirements. Lastly, our study underscores the importance of algorithmic transparency. Recruiters often make incorrect assumptions about how candidate rankings are generated, and a lack of transparency can lead to biased decision-making. For instance, if a recruiter assumes that rankings are based on years of experience, they might prematurely stop exploring the candidate pool when looking for experienced candidates.

5.3. Limitations

To reduce cognitive load and isolate relevant decision-making behaviors, we used simplified candidate profiles and excluded outliers or clearly mismatched applicants. Real CVs often include richer and more diverse information, such as motivation, extracurriculars, and career gaps, which may influence recruiter judgments in practice. Additionally, task features like salary, contract type, and work modality were not included. These simplifications improve experimental control but limit ecological validity.

The quantitative study focuses on shortlisting decisions by crowd workers, who may lack previous recruitment experience and understanding of the positions they are shortlisting for, and may therefore over-rely on specific keywords. Participants also have fewer incentives to reward cultural fitness; real-world recruitment focuses on employer satisfaction and exerts additional pressure on recruiters to factor in cultural fitness and plausibly related factors such as ethnicity in hiring decisions (Almeida et al., 2019; Bonelli & Zhu, 2024; Bye et al., 2014; Wolgast et al., 2018). Despite this fact, participants showed a significant preference for European candidates, although in a limited sample, perhaps also because they are European residents. The qualitative study with experienced recruiters may also have been shaped by the' prior exposure of participants to anti-discrimination training and the reliance on the ' professional networks of the authors for recruitment.

While candidate gender is considered and modeled, our experimental design focuses on ethnicity, especially through ranking algorithms and fitness design. Overall, we examine ethnicity-based discrimination through a European lens, which is reflected in both our candidate profile design and the recruitment of participants. We leave evaluations of gender-based discrimination and intersectional analyses for future work.

We have tested extreme ranking manipulations, which result in strongly favoring one group over another. This departs from more sophisticated fairness-aware ranking algorithms that optimize both fairness and candidate fitness, making our manipulation more evident and therefore limiting the generalization of our findings.

Finally, we focus on the shortlisting stage of the recruitment process. Discrimination may arise earlier (e.g., in job ad targeting) or later (e.g., interviews, job offers), and should be addressed across the full recruitment pipeline. Overall, our findings must be understood within the constraints of our experimental design and the scope of this research.

6. Conclusion and future work

We investigated how task-level factors influence position bias (RQ1) and found that higher selection rates and reduced friction in accessing candidate profiles lead to a weaker preference for selecting individuals near the top of the list. This, in turn, may encourage the consideration of more candidates, creating an opportunity to mitigate demographic imbalances. When examining factors that impact demographic balance (RQ2), we found that proxies for "foreignness", such as experience in non-EU companies and education at non-EU institutions, decrease the likelihood of being shortlisted for job positions in Europe. Ranking EU candidates above non-EU candidates amplifies this effect, whereas reversing the ranking mitigates it–though not symmetrically. This undesirable asymmetry is especially large under high position bias. Our qualitative study with recruiters (RQ3) provides insight into their perceptions, assumptions, and strategies when shortlisting candidates. It suggests that the preference for candidates at the top of a list is driven by the constraints of time and attention, commonly experienced as "recruitment fatigue" and that biases against foreign job seekers originate from assumptions on language, regulations, culture, and prestige.

Research on fair ranking algorithms often assumes that higher visibility in a ranking, i.e., appearing in top positions, directly translates to a greater likelihood of being shortlisted. However, our study indicates that this relationship is influenced by multiple factors and cannot be easily modeled. This is not to suggest that fair ranking algorithms are ineffective, rather that recruitment *outcomes* must be monitored as rigorously as the outputs of ranking algorithms. Fairness initiatives are necessary *a fortiori* to counteract assumptions and disadvantages that hinder vulnerable groups downstream of algorithms and remain hidden from typical fair ranking measures. Developing theoretical models capable of predicting and maintaining demographic balance in recruitment outcomes is essential. Our work highlights the critical role of position bias and discrimination attention against vulnerable candidates. Discrimination is multifaceted and complex, and the development of anti-discrimination frameworks should account for broader contextual factors, additional protected attributes, and intersectional analyses of gender and ethnicity, while also considering the nuances of recruiter-system interactions.

CRediT authorship contribution statement

Alessandro Fabris: Writing – original draft, Methodology, Formal analysis, Conceptualization; **Clara Rus:** Writing – original draft, Methodology, Formal analysis, Conceptualization; **Jorge Saldivar:** Writing – original draft, Methodology, Formal analysis, Conceptualization; **Anna Gatzioura:** Writing – original draft, Methodology, Formal analysis, Conceptualization; **Asia J. Biega:** Writing – original draft, Methodology, Formal analysis, Conceptualization; **Carlos Castillo:** Writing – original draft, Methodology, Formal analysis, Conceptualization.

Acknowledgements

This work is supported by the FINDHR project, Horizon Europe grant agreement ID: 101070212 and by the Alexander von Humboldt Foundation.

Data availability

Data is available at https://doi.org/10.5281/zenodo.17780216.

Appendix A. Detailed description of candidate fitness

Design of fitness distribution. The extent to which the experience, education, and skills of each candidate match corresponding fields in job descriptions determines the overall candidate's fitness. The 20 profiles for each low-variability job are built with the same target fitness in mind. In contrast, the candidates for high-variability jobs are divided into high-fitness (8 profiles), low-fitness (8 profiles), and medium-fitness (4 profiles).

Joint fitness-demographic distribution. We impose an even distribution of fitness across protected groups (Bertrand & Mullainathan, 2004). Low-variability jobs have at least four candidates in each intersectional category (female non-European, female European, male non-European, and male European), and all twenty candidates have the same target fitness. For high-variability jobs, there are two members of each intersectional group among high-fitness candidates and two members among low-fitness candidates. Medium-fitness candidates have random demographics. Among non-European candidates, 50 % are recent immigrants and 50 % are long-term residents. Among high-fitness candidates, there are two long-term residents and two recent migrants. High-fitness candidates are designed to match job requirements better than low-fitness ones. To exemplify, for a position requesting a master's degree, 5 years of relevant experience, and 5 key skills, high-fitness candidates may have (on average) the requested master's degree, 4 years of experience in the position, and 4 relevant skills matching the ones in the opening, while low-fitness candidates might list a bachelor's degree, 2 years of experience, and 3 relevant skills.

Controlling confounders. Within each job, candidates' experience can vary by a maximum of 3 years. This should filter out the effect of age-based preferences in human judgments. Education and skills are also "upper-bounded", so no candidate appears extremely over-qualified for a given position.

Appendix B. Qualitative study supplementary materials

B.1. Participant recruitment and compensation

Recruiters were contacted by email, after collecting statements of interest of alumni in an `Algorithmic Hiring` training program for professionals organized by an EU-funded project in which the institutions of some of the authors are partners (project name withheld for double-blind review). No financial compensation was provided to recruiters.

B.2. Qualitative study structure

We conducted semi-structured interviews combined with a think-aloud protocol to explore participants' decision-making processes during a simulated recruitment task. Each session was divided into five parts.

The session began with a brief introduction outlining the study's purpose and ethical considerations, including assurances of anonymity and data confidentiality. Participants provided informed consent and were asked to allow screen and audio recording, with all recordings scheduled for deletion at the end of the study. Cameras were turned off to minimize distractions, and participants shared basic background information, including their experience in recruitment and the approximate number of recruitment processes they had conducted.

In the task introduction, participants were instructed to evaluate candidates for a given job description using a custom interface. They were encouraged to mimic real-world recruitment behavior and verbalize their thoughts throughout.

Next, participants were asked to elaborate on their selection rationale, the perceived quality of top-ranked candidates, and any observed patterns related to candidate competence, background, or demographics. They were also asked to compare the study interface with platforms they use professionally and to reflect on any prior training related to bias or discrimination in recruitment. The questions are reported in Appendix B.3.

In the fourth segment, participants were shown two alternative candidate rankings–one favoring individuals from disadvantaged backgrounds, the other favoring those from privileged groups. They were asked to consider how such rankings might influence hiring decisions and the broader recruitment process.

Finally, the session concluded with participants ceasing screen sharing and recording. The interviewer thanked them for their time and invited any additional feedback or reflections.

B.3. Qualitative study protocol

The interview sessions were conducted remotely via video conferencing. One researcher acted as the primary facilitator (hereafter, the interviewer), while a second researcher served as an observer and note-taker. Both were members of the research team.

The full interview protocol is detailed below. Instructions delivered to participants in the interviewer's own words are presented in regular text, while scripted statements read verbatim are indicated in **bold**. Descriptions of interviewer actions are presented in narrative form.

`Part 1: Greeting`

1. Begin the video call by greeting the participant and asking for their first name
2. Let the participant know that we will not mention the name or organization of the participant in our research report
3. **The experiment consists of looking at candidates for a series of jobs, and selecting one of them for each job, explaining your thought process as you go through the candidates. We will give you a URL soon, and you will be able to see the candidates**
4. Ask if it is OK for the screen and audio of the call to be recorded, with the recording deleted at the end of the study
   If participant agrees, ask participant to turn off camera, otherwise just take notes
5. Start recording if the participant agreed to be recorded
6. Ask the participant, **Q1. Regarding your background and work experience in relation to recruitment:**
   - **Q1.1 How would you describe, in a short phrase, your experience regarding recruitment?**
   - **Q1.2. How many recruitment processes have you been involved in (1, 2, 3+, 10+, 50+)**
   - **Q1.3. Have you ever worked in an HR department or in an HR role? Which one?**
7. Ask the participant to share their screen
8. Send participant the URL of the experiment via chat
9. Ask participant to read the information sheet
10. Ask participant if they have questions

`Part 2: Behavioral observations and think-aloud over the fitness-based rankings`

1. **You will be shown job descriptions and candidates. We will do this for about 10–15 min, you can stop at any time. After that, we will ask you a few questions**
2. Ask participant to click NEXT and start the experiment
3. **Your task is to select the best candidate for the job description displayed below. All candidates already comply with all legal requirements to take the position (e.g., those who need it, already have a work permit). All candidates have at least**

a lower-secondary education diploma. To view more information about a candidate, you can click on the "View" button. After selecting the candidate click the "Next" button to navigate to the next jobs

4. **While you complete this task, to the extent possible, please try to mimic the processes you would use at work. Also, please speak your thoughts out loud to explain your process: including the steps and decisions you're taking**
5. After the participant finishes the first job description (it should take 10–15 min), i.e., selects one candidate and click on next, ask
   - If you have just 10 min more, we go straight to questions, but if you have 20 min more, we can do one more job and then go to the questions, please. As you prefer

`Part 3: Behavioral and experiment questions`

1. Thank the participant and indicate you will now ask some questions to understand their thought process better
   - **Q3.1 Could you please summarize which factors were most important for you when deciding whom to select?**
   - **Q3.2 Did you have any thoughts about the way people were ranked/ordered in the list?**
     [If not mentioned spontaneously] Did you think the list placed "good" candidates at the top?
   - **Q3.3 Did you think that the candidates were different from each other, within each of the jobs?**
     Q3.3.1 [If yes] In what ways?
   - **Q3.4 Did you think that the candidates were different from each other in terms of their background or demographics?**
     Q3.4.1. [If yes] Could you name the background or demographic characteristics you thought were different among the candidates?
     Q3.4.2 How did you identify or infer those characteristics?
     Q3.4.3 What impact could these characteristics have in a real recruitment process?
   - **Q3.5 In which ways was this system/interface different from systems/interfaces you have used for a similar task in recruitment?**
   - **Q3.6 Did you receive any training on non-discrimination or awareness of biases in recruitment?**

`Part 4: Comment on discriminatory and positive action rankings`

1. **As a final step in this study, we would like you to assess two other job candidate rankings, one where candidates from a disadvantaged group are displayed at the top, and another where candidates from a privileged group are displayed at the top**
2. Copy-paste links of the rankings into the chat window
3. **The idea is to imagine that a recruiter was to use one of these rankings. Could you share your thoughts on if and how each of these two rankings might impact the recruitment process?**

`Part 5: Closing`

1. Tell the participant they can stop sharing their screen
2. Thank the person for their time and for participating in this experiment
3. Stop recording, if recording
4. Let the participant know we will publish (anonymized) results from this study in the coming months–we will not mention their name or organization
5. Let the participant know we will conduct further studies, and we will not send more than one invitation per month
6. Ask if there are any further questions or comments
7. Thank the participant again
8. Close the videoconference session

B.4. Qualitative data analysis and codes

The analytical process proceeded as follows: First, the research team established a set of initial codes through iterative discussion, grounded in the study's research questions. Two authors (hereafter, coders) independently applied this coding scheme to three of the twelve interview transcripts, assigning excerpts to the relevant codes. Following this initial round, the coders engaged in a consensus-building process to resolve discrepancies and refine code definitions, resulting in a shared understanding of the application of the coding schema. Subsequently, the remaining nine transcripts were divided between the two coders, each analyzing a subset independently. After all transcripts were coded, we compiled the coded excerpts by category and further organized them by participant. This structure enabled a systematic synthesis of thematic insights within each code.

The final phase of analysis involved reviewing the aggregated coded data to derive high-level patterns and interpretations. Each code corresponded to a specific analytical focus:

- **Assumptions** reflected the implicit or explicit criteria participants employed when determining candidate suitability.
- **Ranking Perception** captured participant responses to the system-generated candidate orderings and their perceptions of its validity.
- **Candidate Representation** addressed participant observations about demographic diversity and its visibility in the candidate lists.

- **Experimental Fidelity** included commentary on the perceived realism of the experimental setup in relation to actual recruitment practices and platforms.
- **Positive action** aggregated reflections on the influence of positive action measures and fairness-oriented rankings on decision-making.

References

Agarwal, A., Wang, X., Li, C., Bendersky, M., & Najork, M. (2019). Addressing trust bias for unbiased learning-to-rank. In L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J.McAuley, R. Baeza-Yates, & L. Zia (Eds.), *The world wide web conference, WWW 2019, San Francisco, CA, Usa, May 13–17, 2019* (pp. 4–14). ACM. https://doi.org/10.1145/3308558.3313697

Ainsworth, S., & Pekarek, A. (2022). Gender in human resources: Hiding in plain sight. *Human Resource Management Journal*, *32*(4), 890–905. 10.1111/1748-8583.12437.

Almeida, S., Fernando, M., Hannif, Z., & Dharmage, S. C. (2015). Fitting the mould: The role of employer perceptions in immigrant recruitment decision-making. *The International Journal of Human Resource Management*, *26*(22), 2811–2832. https://doi.org/10.1080/09585192.2014.1003087.

Almeida, S., Waxin, M.-F., & Paradies, Y. (2019). Cultural capital of recruitment decision-makers and its influence on their perception of person-organisation fit of skilled migrants. *International Migration*, *57*(1), 318–334. https://doi.org/10.1111/imig.12539.

Alvarez, J. M., Mastropietro, A., & Ruggieri, S. (2025). The initial screening order problem. In *Proceedings of the eighteenth ACM international conference on web search and data mining* (pp. 165–174).

Amigó, E., Deldjoo, Y., Mizzaro, S., & Bellogín, A. (2023). A unifying and general account of fairness measurement in recommender systems. *Information Processing & Management*, *60*(1), 103115. https://doi.org/10.1016/j.ipm.2022.103115

Baert, S., Cockx, B., Gheyle, N., & Vandamme, C. (2015). Is there less discrimination in occupations where recruitment is difficult? *ILR Review*, *68*(3), 467–500. https://doi.org/10.1177/0019793915570873.

Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org/.

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, *106*(6), 1437–1475. https://doi.org/10.1257/aer.20140571

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. *Handbook of Economic Field Experiments*, *1*, 309–393. https://doi.org/10.1016/bs.hefe.2016.08.004.

Bertrand, M., & Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013.

Biega, A. J., Gummadi, K. P., & Weikum, G. (2018). Equity of attention: Amortizing individual fairness in rankings. In K. Collins-Thompson, Q. Mei, B. D.Davison, Y. Liu, & E. Yilmaz (Eds.), *The 41st international ACM SIGIR conference on research & development in information retrieval, SIGIR 2018, ANN Arbor, MI, USA, July 08–12, 2018* (pp. 405–414). ACM. https://doi.org/10.1145/3209978.3210063

Bigdeli, A., Arabzadeh, N., Seyedsalehi, S., Zihayat, M., & Bagheri, E. (2022). Gender fairness in information retrieval systems. In E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S.Culpepper, & G. Kazai (Eds.), *SIGIR '22: The 45th international ACM SIGIR conference on research and development in information retrieval, Madrid, Spain, July 11, - 15, 2022* (pp. 3436–3439). ACM. https://doi.org/10.1145/3477495.3532680

Bingham, A. J., & Witkowsky, P. (2021). Deductive and inductive approaches to qualitative data analysis. *Analyzing and Interpreting Qualitative Data: After the Interview*, *1*, 133–146.

Bonelli, N., & Zhu, H. (2024). The myth of cultural fit in recruitment job interviews. *World Englishes*. https://doi.org/10.1111/weng.12710.

Bye, H. H., Horverak, J. G., Sandal, G. M., Sam, D. L., & Van de Vijver, F. J. R. (2014). Cultural fit and ethnic background in the job interview. *International Journal of Cross Cultural Management*, *14*(1), 7–26. https://doi.org/10.1177/1470595813491237.

Carterette, B. (2011). System effectiveness, user models, and user utility: a conceptual framework for investigation. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 903–912). https://doi.org/10.1145/2009916.2010037.

Chen, L., Ma, R., Hannák, A., & Wilson, C. (2018). Investigating the impact of gender on rank in resume search engines. In R. L. Mandryk, M. Hancock, M. Perry, & A. L.Cox (Eds.), *Proc. of the 2018 CHI conf. on human factors in computing systems, CHI 2018, Montreal, QC, Canada, April 21–26, 2018* (p. 651). ACM. https://doi.org/10.1145/3173574.3174225

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 87–94).

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/. URL visited on 28th August 2025.

De-Arteaga, M., Romanov, A., Wallach, H. M., Chayes, J. T., Borgs, C., Chouldechova, A., Geyik, S. C., Kenthapadi, K., & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In d. boyd, & J. H. Morgenstern (Eds.), *Proc. of the conf. on fairness, accountability, and transparency, FAT* 2019, Atlanta, GA, USA, January 29–31, 2019* (pp. 120–128). ACM. https://doi.org/10.1145/3287560.3287572

Derous, E., Nguyen, H.-H., & Ryan, A. M. (2009). Hiring discrimination against arab minorities: Interactions between prejudice and job characteristics. *Human Performance*, *22*(4), 297–320. https://doi.org/10.1080/08959280903120261.

Derous, E., & Ryan, A. M. (2019). When your resume is (not) turning you down: Modelling ethnic bias in resume screening. *Human Resource Management Journal*, *29*(2), 113–130. https://doi.org/10.1111/1748-8583.12217.

Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., & Carterette, B. (2020). Evaluating stochastic rankings with expected exposure. In M. d'Aquin, S. Dietze, C. Hauff, E. Curry, & P. Cudré-Mauroux (Eds.), *CIKM '20: The 29th ACM international conference on information and knowledge management, virtual event, Ireland, October 19–23, 2020* (pp. 275–284). ACM. https://doi.org/10.1145/3340531.3411962

Dineen, B. R., Noe, R. A., & Wang, C. (2004). Perceived fairness of web-based applicant screening procedures: Weighing the rules of justice and the role of individual differences. *Human resource management: published in cooperation with the school of business administration, the university of michigan and in alliance with the society of human resources management*, *43*(2-3), 127–145. https://doi.org/10.1002/hrm.20011.

Ekstrand, M. D., Das, A., Burke, R., Diaz, F. et al. (2022). Fairness in information access systems. *Foundations and Trends® in Information Retrieval*, *16*(1-2), 1–177. 10.1561/1500000079.

European Union (2024). Artificial intelligence act. Regulation (EU) 2024/1689 of the European Parliament and of the Council. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

Eurostat (2011). Statistics in focus: 6.5 % of the EU population are foreigners and 9.4 % are born abroad - issue number 34/2011. https://ec.europa.eu/eurostat/web/products-statistics-in-focus/-/ks-sf-11-034. URL visited on 30th September 2023.

Eurostat (2023). Data browser: First permits by reason, length of validity and citizenship. https://ec.europa.eu/eurostat/databrowser/view/migr_resfirst/default/table?lang=en. URL visited on 30th September 2023.

Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., Zuiderveen Borgesius, F., & Biega, A. J. (2025). Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3696457.

Fazelpour, S., & De-Arteaga, M. (2022). Diversity in sociotechnical machine learning systems. *Big Data & Society*, *9*(1), 20539517221082027. https://doi.org/10.1177/20539517221082027.

Fuller, J., Raman, M., Sage-Gavin, E., & Hines, K. (2021). Hidden workers: Untapped talent. Technical Report Harvard Business School. https://www.hbs.edu/managing-the-future-of-work/Documents/research/hiddenworkers09032021.pdf.

Geyik, S. C., Ambler, S., & Kenthapadi, K. (2019). Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2221–2231). https://doi.org/10.1145/3292500.3330691.

Geyik, S. C., Guo, Q., Hu, B., Ozcaglar, C., Thakkar, K., Wu, X., & Kenthapadi, K. (2018). Talent search and recommendation systems at linkedin: Practical challenges and lessons learned. In K. Collins-Thompson, Q. Mei, B. D.Davison, Y. Liu, & E. Yilmaz (Eds.), *The 41st international ACM SIGIR conf. on research & development in information retrieval, SIGIR 2018, ANN Arbor, MI, USA, July 08–12, 2018* (pp. 1353–1354). ACM. https://doi.org/10.1145/3209978.3210205

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089.

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 478–479). https://doi.org/10.1145/1008992.1009079.

Groves, L., Metcalf, J., Kennedy, A., Vecchione, B., & Strait, A. (2024). Auditing work: Exploring the new york city algorithmic bias audit regime. In *The 2024 ACM conference on fairness, accountability, and transparency, FACCT 2024, Rio de Janeiro, Brazil, June 3–6, 2024* (pp. 1107–1120). ACM. https://doi.org/10.1145/3630106.3658959

Hannak, A., Wagner, C., García, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In C. P. Lee, S. E. Poltrock, L. Barkhuus, M. Borges, & W. A.Kellogg (Eds.), *Proc. of the 2017 ACM conf. on computer supported cooperative work and social computing, CSCW 2017, Portland, OR, USA, February 25, - March 1, 2017* (pp. 1914–1933). ACM. https://doi.org/10.1145/2998181.2998327

Hireview (2022). Explainability statement. https://hirevue-api.dev-directory.com/wp-content/uploads/2022/04/HV_AI_Short-Form_Explainability_1pager.pdf. URL visited on 28th August 2025.

Hsu, J. (2020). Can AI hiring systems be made antiracist? makers and users of AI-assisted recruiting software reexamine the tools' development and how they're used-[news]. *IEEE Spectrum*, *57*(9), 9–11. https://doi.org/10.1109/MSPEC.2020.9173891.

Huang, Y., Liu, D.-R., & Lee, S.-J. (2023). Talent recommendation based on attentive deep neural network and implicit relationships of resumes. *Information Processing & Management*, *60*(4), 103357. https://doi.org/10.1016/j.ipm.2023.103357

Islam, M. A., Srikant, R., & Basu, S. (2019). Micro-browsing models for search snippets. In *35th IEEE international conference on data engineering, ICDE 2019, Macao, China, April 8–11, 2019* (pp. 1904–1909). IEEE. https://doi.org/10.1109/ICDE.2019.00206

Kang, S. K., DeCelles, K. A., Tilcsik, A., & Jun, S. (2016). Whitened résumés: Race and self-presentation in the labor market. *Administrative Science Quarterly*, *61*(3), 469–502. https://doi.org/10.1177/0001839216639577.

Kappen, M., & Naber, M. (2021). Objective and bias-free measures of candidate motivation during job applications. *Scientific Reports*, *11*(1), 21254. https://doi.org/10.1038/s41598-021-00659-y.

Kaya, M., & Bogers, T. (2023). Understanding recruiters' information seeking behavior in talent search. In *Proceedings of the 2023 conference on human information interaction and retrieval* (pp. 14–23). https://doi.org/10.1145/3576840.3578311.

Kelan, E. K. (2024). Algorithmic inclusion: Shaping the predictive algorithms of artificial intelligence in hiring. *Human Resource Management Journal*, *34*(3), 694–707. https://doi.org/10.1111/1748-8583.12511.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology*, *100*(1), 128. https://doi.org/10.1037/a0036734.

Lahey, J., & Oxley, D. R. (2018). Discrimination at the intersection of age, race, and gender: Evidence from a lab-in-the-field experiment. Technical Report National Bureau of Economic Research Cambridge, MA, USA. https://doi.org/10.1002/pam.22281.

Lavanchy, M., Reichert, P., Narayanan, J., & Savani, K. (2023). Applicants' fairness perceptions of algorithm-driven hiring procedures. *Journal of Business Ethics*, *188*(1), 125–150. https://doi.org/10.1007/s10551-022-05320-w.

Leony, R. D., Mataheru, S., Sirait, K. S., & Prasandy, T. (2024). Evaluation analysis of applicant tracking system (ATS) implementation in companies and recruitment agency. In *2024 Ninth international conference on informatics and computing (ICIC)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICIC64337.2024.10957063.

Lewis, G., & Mohapatra, M. (2023). The most in-demand jobs on linkedin right now. https://www.linkedin.com/business/talent/blog/talent-strategy/most-in-demand-jobs. URL visited on 30th September 2023.

Li, D., Raymond, L. R., & Bergman, P. (2020). Hiring as exploration. Technical Report National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w27736/w27736.pdf.

Maćkowiak, B., Matějka, F., & Wiederholt, M. (2023). Rational inattention: A review. *Journal of Economic Literature*, *61*(1), 226–273. https://doi.org/10.1257/jel.20211524

Miller, A. (2018). Want less-biased decisions? Use algorithms. https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms. URL visited on 28th August 2025.

Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, *27*(1), 2:1–2:27. https://doi.org/10.1145/1416950.1416952

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature*, *56*(3), 799–866. https://doi.org/10.1257/jel.20161309

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008.

Peng, A., Nushi, B., Kıcıman, E., Inkpen, K., Suri, S., & Kamar, E. (2019). What you see is what you get? The impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI conference on human computation and crowdsourcing* (pp. 125–134). (vol. *7*). https://doi.org/10.1609/hcomp.v7i1.5281.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 469–481). https://doi.org/10.1145/3351095.3372828.

Rakova, B. (2023). Speculative f(r)iction in generative AI. https://foundation.mozilla.org/en/blog/speculative-friction-in-generative-ai/. URL visited on 30th September 2024.

Rigotti, C., & Fosch-Villaronga, E. (2024). Fairness, AI & recruitment. *Computer Law & Security Review*, *53*, 105966. https://doi.org/10.1016/j.clsr.2024.105966

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, *33*(4), 294–304. https://doi.org/10.1108/eb026647

Russell-Rose, T., & Chamberlain, J. (2016). Searching for talent: The information retrieval challenges of recruitment professionals. *Business Information Review*, *33*(1), 40–48. https://doi.org/10.1177/0266382116631849

Schumann, C., Lang, Z., Foster, J., & Dickerson, J. (2019). Making the cut: A bandit-based approach to tiered interviewing. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/file/d3fad7d3634dbfb61018813546edbccb-Paper.pdf.

Seppälä, P., & Małecka, M. (2024). Ai and discriminative decisions in recruitment: Challenging the core assumptions. *Big Data & Society*, *11*(1), 20539517241235872. https://doi.org/10.1177/20539517241235872.

Simon, V., Rabin, N., & Gal, H. C.-B. (2023). Utilizing data driven methods to identify gender bias in linkedin profiles. *Information Processing & Management*, *60*(5), 103423. https://doi.org/10.1016/j.ipm.2023.103423

Singh, A., & Joachims, T. (2018). Fairness of exposure in rankings. In Y. Guo, & F. Farooq (Eds.), *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2018, London, UK, August 19–23, 2018* (pp. 2219–2228). ACM. https://doi.org/10.1145/3219819.3220088

Steinpreis, R. E., Anders, K. A., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles*, *41*(7), 509–528. https://doi.org/10.1023/A:1018839203698.

Sühr, T., Hilgard, S., & Lakkaraju, H. (2021). Does fair ranking improve minority outcomes? Understanding the interplay of human and algorithmic biases in online hiring. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 989–999). https://doi.org/10.1145/3461702.3462602.

Team, I. E. (2025). What are resume keywords and why are they important? https://ca.indeed.com/career-advice/resumes-cover-letters/resume-keywords. Online; accessed 8 April 2025.

Vaishampayan, S., Farzanehpour, S., & Brown, C. (2023). Procedural justice and fairness in automated resume parsers for tech hiring: Insights from candidate perspectives. In *IEEE symposium on visual languages and human-centric computing, VL/HCC 2023, Washington, DC, USA, October 3–6, 2023* (pp. 103–108). IEEE. https://doi.org/10.1109/VL-HCC57772.2023.00019

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. In M. C. Elish, W. Isaac, & R. S.Zemel (Eds.), *Facct '21: 2021 ACM conf. on fairness, accountability, and transparency, virtual event / Toronto, Canada, March 3–10, 2021* (pp. 666–677). ACM. https://doi.org/10.1145/3442188.3445928

Wolgast, S., Björklund, F., & Bäckström, M. (2018). Applicant ethnicity affects which questions are asked in a job interview. *Journal of Personnel Psychology*. https://psycnet.apa.org/doi/10.1027/1866-5888/a000197.

Yang, K., & Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In *Proc. of the 29th international conference on scientific and statistical database management* SSDBM '17. https://doi.org/10.1145/3085504.3085526.

Yarger, L., Cobb Payton, F., & Neupane, B. (2020). Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Information Review*, *44*(2), 383–395.

Zehlike, M., Yang, K., & Stoyanovich, J. (2022a). Fairness in ranking, Part I: Score-based ranking. *ACM Computing Surveys*, *55*(6), 1–36. https://doi.org/10.1145/3533379.

Zehlike, M., Yang, K., & Stoyanovich, J. (2022b). Fairness in ranking, part II: Learning-to-rank and recommender systems. *ACM Computing Surveys*, *55*(6), 1–41. https://doi.org/10.1145/3533380.