Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey

ALESSANDRO FABRIS, Max Planck Institute for Security and Privacy, Bochum, Germany
NINA BARANOWSKA, Radboud University, Nijmegen, The Netherlands
MATTHEW J. DENNIS, Eindhoven University of Technology, Eindhoven, The Netherlands
DAVID GRAUS, Randstad, Diemen, The Netherlands
PHILIPP HACKER, European University Viadrina, Frankfurt (Oder), Germany
JORGE SALDIVAR, Universitat Pompeu Fabra, Barcelona, Spain
FREDERIK ZUIDERVEEN BORGESIUS, Radboud University, Nijmegen, The Netherlands
ASIA J. BIEGA, Max Planck Institute for Security and Privacy, Bochum, Germany

Employers are adopting algorithmic hiring technology throughout the recruitment pipeline. Algorithmic fairness is especially applicable in this domain due to its high stakes and structural inequalities. Unfortunately, most work in this space provides partial treatment, often constrained by two competing narratives, optimistically focused on replacing biased recruiter decisions or pessimistically pointing to the automation of discrimination. Whether, and more importantly *what types of*, algorithmic hiring can be less biased and more beneficial to society than low-tech alternatives currently remains unanswered, to the detriment of trustworthiness. This multidisciplinary survey caters to practitioners and researchers with a balanced and integrated coverage of systems, biases, measures, mitigation strategies, datasets, and legal aspects of algorithmic hiring and fairness. Our work supports a contextualized understanding and governance of this technology by highlighting current opportunities and limitations, providing recommendations for future work to ensure shared benefits for all stakeholders.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Information systems** → *Data management systems*; • **Social and professional topics** → **Computing / technology policy**;

Additional Key Words and Phrases: Algorithmic hiring, Online recruitment, Algorithmic fairness, Bias, Anti-discrimination

Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, and Frederik Zuiderveen Borgesius are listed in alphabetical order.

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

Authors' Contact Information: Alessandro Fabris (corresponding author), Max Planck Institute for Security and Privacy, Bochum, Germany; e-mail: alessandro.fabris@mpi-sp.org; Nina Baranowska, Radboud University, Nijmegen, The Netherlands; e-mail: nina.baranowska@ru.nl; Matthew J. Dennis, Eindhoven University of Technology, Eindhoven, The Netherlands; e-mail: m.j.dennis@tue.nl; David Graus, Randstad, Diemen, The Netherlands; e-mail: david@graus.nu; Philipp Hacker, European University Viadrina, Frankfurt (Oder), Germany; e-mail: hacker@europa-uni.de; Jorge Saldivar, Universitat Pompeu Fabra, Barcelona, Spain; e-mail: jorge.saldivar@upf.edu; Frederik Zuiderveen Borgesius, Radboud University, Nijmegen, The Netherlands; e-mail: frederikzb@cs.ru.nl; Asia J. Biega, Max Planck Institute for Security and Privacy, Bochum, Germany; e-mail: asia.biega@mpi-sp.org.

ACM 2157-6912/2025/1-ART16
https://doi.org/10.1145/3696457

ACM Reference format:
Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. 2025. Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey. *ACM Trans. Intell. Syst. Technol.* 16, 1, Article 16 (January 2025), 54 pages.
https://doi.org/10.1145/3696457

1 Introduction

New algorithms for **Human Resources (HR)** are developed and deployed every year. By one count, there are over 250 AI tools for HR on the market [99], with entire manuals for HR professionals available on the topic [80]. The average job posting yields more than 100 candidates [101, 206]. These mutually reinforcing factors were accelerated by the COVID-19 pandemic and recent advances in AI [170, 207]. As a result, prospective job applicants and their chances of success are increasingly influenced by algorithmic hiring technology, including automated job descriptions, resume parsers, and video interviews.

Workplaces and labor markets are fraught with biases, imbalances, and patterns of discrimination against vulnerable groups, including women, ethnic minorities, and people with disabilities [15, 23, 215]. While algorithmic hiring represents an opportunity to mitigate these biases, it also runs the risk of reinforcing and amplifying them, causing harm and hindering trustworthiness [168]. The debate on this topic is often polarized between techno-enthusiasm [139, 177] and pessimism [9, 132] due to partial perspectives on a field that is large and complex. In this article, we offer a multidisciplinary survey of fairness and bias in algorithmic hiring centered on computer science (focused on systems, algorithms, metrics, and datasets) and informed by related disciplines. We critically analyze available resources and methods, highlight common challenges, and identify opportunities to advance the field.

Related Work. Bogen and Rieke [30] presented a technical report that described algorithmic hiring tools available in 2018, together with selected sources of bias, and provided a US-centric review of relevant laws and policies. Köchling and Wehner [149] conducted a review of algorithms in HR from a business research perspective, focused on non-empirical articles. Their qualitative discussion raises awareness on discrimination in HR algorithms without delving into measures and mitigation strategies. The prior work most closely aligned to ours is the work of Rieskamp et al. [217], surveying nine articles on bias mitigation for algorithmic hiring. Our survey expands on this work by presenting more bias mitigation techniques, covering fairness measures, describing available datasets, and by situating them in the broader social and legal context characterizing algorithmic hiring. In concurrent work, Kumar et al. [154] survey the literature on fair recommender systems in the recruitment domain. While similar in spirit, their work focuses on ranking; our work presents a broader view of algorithmic hiring across different tasks and hiring stages, covering many technologies and **Bias Conducive Factors (BCFs)** throughout the algorithmic hiring pipeline.

Contributions and Audience. This work provides a contextualized treatment of fairness and bias in algorithmic hiring. It was carried out by a team with mixed backgrounds in computer science, law, and philosophy, including practitioners developing algorithmic hiring products. Our contributions, catering especially to practitioners and researchers, can roughly be divided as follows. Practitioners such as data scientists, engineers, and product managers will find (1) a detailed list and description of domain-specific factors that lead to biases in their systems (Section 3), (2) methods to mitigate these biases (Section 5), (3) guidance on their applicability in practice (Section 7), and (4) pointers to key legal references in the EU and the US (Section 8). Researchers will benefit from (4) an up-to-date description of hiring technology (Section 2), (5) a unified treatment of fair hiring measures (Section 4), and (6) a collection of datasets in this space (Section 6). Overall, an integrated coverage

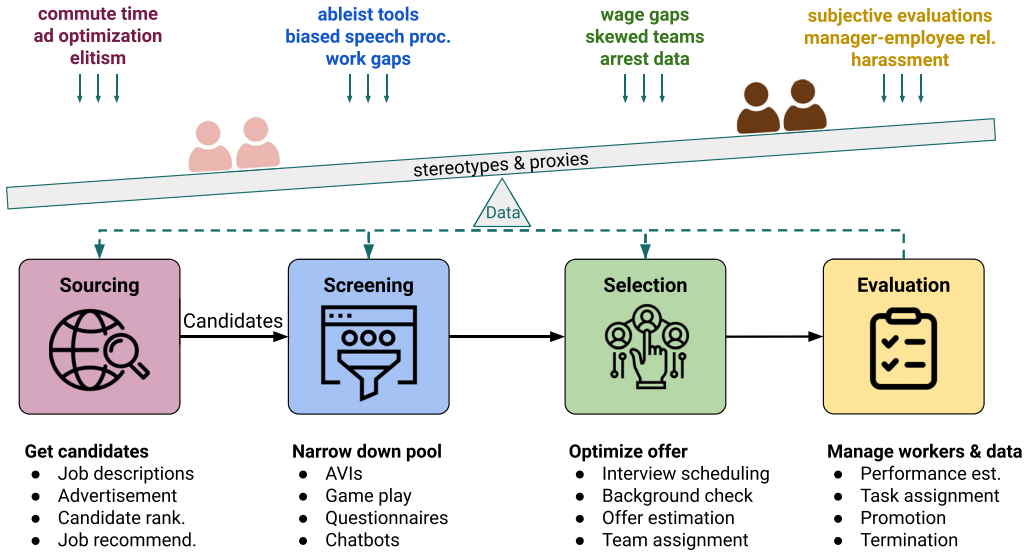

Fig. 1. Stages of algorithmic hiring with the main tools (below) and BCFs (above) tilting data against vulnerable groups, jointly with stereotypes and sensitive attribute proxies.

of these topics provides (7) a gentle primer for readers who are not experts in the field and (8) highlights important gaps and promising directions for future work in computer science at the intersection with law and policy (Section 10). Considering a broader audience, Sections 2, 3, 8, 10, and 11 cover a shared background which should be relevant for all readers, including HR professionals and legal scholars.

Structure. The remainder of this work is organized as follows. Section 2 introduces the main stages and systems in the algorithmic hiring pipeline. Section 3 focuses on bias, summarizing the most important factors in the labor market, the recruitment sector, and the tech industry that can lead to unfair recruitment systems. Sections 4–6 result from a systematic review of the literature with methods described in Appendix A. Sections 4 and 5 describe the main fairness measures and mitigation approaches, while Section 6 presents the datasets used in the algorithmic hiring literature. Section 7 discusses practical aspects of anti-discrimination in algorithmic hiring guiding the choice of fairness measures and mitigation strategies. Sections 8 and 9 widen the perspective beyond computer science by outlining the main legal frameworks and situating algorithmic hiring in its broader socio-technical context. Section 10 summarizes opportunities, limitations, and recommendations for future work; Section 11 provides concluding remarks.

2 The Algorithmic Hiring Pipeline

Algorithmic hiring comprises algorithms, tools, and systems to automate or assist HR decisions on candidate recruitment and evaluation. Elaborating on previous work from civil society [30], regulatory bodies [84], and academia [209], we distinguish four stages in the algorithmic hiring pipeline, reported in Figure 1.

Sourcing. The first stage of the hiring pipeline provides employers with a large pool of candidates for a given position. Historically, this is the hiring stage where algorithms are most prominent and well-researched. The main technological solutions and systems are summarized below.

— *Job advertisement* consists of tools to describe a vacancy and make it visible. Job descriptions [128], possibly written by language models [207], are optimized and shared through suitable channels, including employer websites [43], dedicated platforms [198], and ad delivery services [4, 158].
— *Search, ranking, and recommendation* algorithms favor matches between jobs and job seekers by ranking candidates for openings [103] and vice versa: recommending openings to job seekers [185], based on a combination of machine learning, Boolean matching, and sorting (e.g., skills, location, industry), supported by information extraction systems [47]. Fairness is a critical and well-studied property for these systems [154, 265].
— *Social networks*, especially those embedded in job platforms, play an important role in the visibility of candidates [239] and their awareness of opportunities [96], favoring professional connections [39], contributing to the dissemination of information, and facilitating referral [284].

It is worth noting that stage categorization is elastic and usage-dependent. For example, algorithms that extract information from CVs to score employees against job descriptions are central to the sourcing stage, but may also be used for screening [58].

Screening. After sourcing, employers must narrow down large pools of candidates into manageable subsets that can be more thoroughly evaluated by HR specialists and other employees.

— *Gameplay* is used for psychometric assessment and soft skills measurement. Companies develop proprietary suites of games and make them available to candidates, e.g., through a mobile app. Models analyze their gameplay to explicitly estimate competency scores [189] or implicitly look for similarities with desirable candidates such as current employees [271].
— ***Asynchronous Video Interviews (AVI)*** are recordings of candidates' answers to a specific set of questions in front of a camera. AVI models rely on different data modalities, including visual (e.g., facial expressions), verbal (e.g., length of sentences), and paraverbal features (e.g., tone) to infer a variety of traits related to personality and hireability [33, 120].
— *Questionnaires* in hiring cover a wide range of purposes, such as personality assessment [129], job performance inference [38, 95], or explicitly asking information about job requirements to filter candidates (e.g., Driver's license ownership) or prioritize them (e.g., years of experience).
— *Chatbots* can mediate interactions between employers and employees by asking basic screening questions for job seekers and scheduling interviews [30, 44]. Advancements in large language models [190] are likely to increase the influence of this technology on hiring processes, broadening its reach to other stages of recruitment, including sourcing and evaluation [170, 224].

Selection. After screening, candidates are interviewed by HR specialists or other employees. These interviews can include technical and cultural questions influenced by previous stages, e.g., by questionnaire answers. Several tools are available at this stage to help employers select the most desirable job seekers and extend them a suitable offer.

— *Compensation and benefit optimization* allows employers to target their offers to each candidate, thanks to tools that estimate the likelihood of acceptance based on salary, bonus, stock options, and other benefits [30, 176].
— *Background checks*, primarily focused on criminal records [59, 66, 157], are used by employers to obtain additional information on candidates before hiring them. Background checks can also target social media [127].

— *Team placement* concerns the assignment of a selected candidate to a specific role within a team or project [107, 202].

Evaluation. The hired employees are managed and evaluated on their career progression. While, strictly speaking, these are post-hiring processes, it is fundamental to consider them due to their feedback loops on algorithmic recruitment.

— *Performance and career management* tools facilitate employee development and monitoring. Technology in this area supports training assignment [36], work allocation [253], career development [188], along with monitoring and estimation through increasingly available tracking technology [64, 140, 232, 236].
— *Engagement analytics* measures employee satisfaction, commitment, and retention probability, often leveraging dedicated questionnaires and surveys [149, 277].
— *Career progression* deals with promotion and dismissal of employees. Succession plans and promotions can be partially automated [169, 188]. Turnover prediction is an active area of research, typically presented as part of analytics to favor retention [204]. It is worth noting that turnover data, frequently used to improve hiring systems [38], can also be repurposed toward employee termination. Indeed 98% of 300 HR leaders from US companies interviewed in a 2023 survey, report that software and algorithms will assist them with layoff decisions throughout the year [259].

The progression of candidates through the stages of the hiring pipeline is accompanied by a feedback loop in which the decisions at each stage generate data that influence the remaining stages in subsequent interactions. Evaluation, for example, can lead to job termination impacting tenure, which, in turn, represents a key desideratum and prediction target for sourcing and screening algorithms. These data are influenced by a diverse and complex set of factors, many of which can lead to undesirable discrimination, as described in the next section.

3 BCFs

Hiring data and algorithms can display undesirable groupwise patterns caused by BCFs affecting the employment domain. These patterns put some job seekers at a systematic disadvantage based on sensitive attributes such as age, disability, gender, religion or belief, racial or ethnic origin, or sexual orientation. Overall, disparities may exist in data sample composition and feature values across sensitive groups that reflect and amplify problematic structural differences in society. First, hiring datasets display measurement errors whose severity varies with protected groups, most critically in target variables related to employability, reflecting current biases in human ratings. Moreover, biased decisions at early stages result in downstream samples misrepresenting certain groups. Finally, there are some higher-order effects caused by technological blind spots and biases in external tools integrated into the hiring pipeline. Some of these biases are reported in Figure 1; at their root there are two overarching BCFs that are worth highlighting from the outset.

Stereotypes are widely held beliefs about groups of individuals with a common trait, including their propensity and ability to perform a given job [118]. Stereotypes are sustained by culture, socialization, and experience [50, 50]. They are often acquired at an early age [134] and can be activated unconsciously [181]. Even when outspokenly rejected, stereotypes affect the lives of individuals both descriptively and prescriptively based on coarse categories [76]. In turn, this ends up shaping expectations about the qualities, priorities, and needs that people have about themselves and others, including, and perhaps especially, about work [29, 50, 118, 248]. For example, *agency* (orientation toward leadership and goal attainment) is stereotypically associated with men, while communion (warmth and propensity for care) is frequently associated with women [72, 118], with far-reaching effects of gender roles and expectations in employment [118, 215].

Sensitive Attribute Proxies. Stereotype activation and entrenchment is not always direct; stereotypes about a group can be activated by proxies [211]. This is especially true for algorithms trained to make inferences from inputs that are strongly correlated with sensitive attributes. For instance, video interviews and resumes contain a wealth of information on gender, race, and other sensitive attributes [63, 69, 189, 218, 222], which can lead algorithms to learn stereotypical associations between sensitive and target variables encoded in the data. More specifically, voice timbre and physical appearance in videos may be used as proxies for gender, while names and spoken languages in CVs may correlate heavily with certain migration backgrounds. Sensitive attribute proxies allow models to learn and reflect the diverse BCFs described in the following section.

3.1 Institutional Biases

The first family of BCFs we introduce are institutional biases. These are practices, habits, and norms shared at institutions, such as companies and societies, which reflect negatively on the probability of positive algorithmic hiring outcomes for disadvantaged groups.

Direct discrimination takes place when disparate outcomes are explicitly caused by sensitive attributes. This type of bias can be difficult to prove, as it requires careful control over non-sensitive attributes to keep them as constant as possible, while only varying sensitive attributes. Famous correspondence studies have shown that black applicants are less likely to be contacted after applying for a job compared to otherwise identical white candidates [23, 147]. Similar results have been found in interview evaluations provided by US-based judges, showing a preference for standard American accents over international ones [67, 164, 205]. Numerous field experiments using matched pairs of applicants have highlighted discrimination against women and non-whites; experiments also point to a risk of direct discrimination against disabled and older applicants [208, 215]. This is the most obvious BCF; a variety of more subtle ones are listed below.

Horizontal Segregation. Job segregation (i.e., the world today) plays a fundamental role in hiring decisions (i.e., the world tomorrow). Prior experience is considered a fundamental predictor of suitability for a given position [101]. Horizontal segregation concerns differences in employment rate across industry sectors associated with sensitive attributes, such as gender and race [28, 243]. Strong gender patterns in diverse regions of the world are linked to persistent gender stereotypes about agency and communion that shape our perception and expectations [42, 83, 109]. Most field experiments, for example, demonstrate discrimination against women in men-dominated jobs and vice versa [215, 216].

Vertical segregation summarizes differences in career progression to leadership positions. Predominantly analyzed with respect to binary gender [53, 82], recent studies have also highlighted *glass ceiling effects* for non-binary workers [62], racial minorities [117], and intersectional identities [273]. When translated into data and dignified with a *ground truth* status, vertical segregation leads to models that reinforce wage gaps [222] and lack of diversity in high-status positions.

Cultural fit is often considered predictive of a candidate's ability to conform and adapt to the core values and collective behaviors of an organization. Evaluations of cultural fit by recruiters, subjective or objective, can contribute to maintaining a uniform workforce in the company, especially in more senior positions, reinforcing horizontal and vertical segregation [3, 219].

Elitism in recruiter evaluations favors candidates educated at prestigious institutions who can also list specific extracurricular accomplishments correlated with socioeconomic status [218], further entrenching family status and social class [51, 184].

Biased employee evaluation has different forms and derives from multiple causes. Two key drivers are stereotypes and employee–manager relationships. Gender and race stereotypes about competence drive people's perception on the workplace and potentially bias supervisor evaluations against female and black employees [116, 229, 233, 242]. Furthermore, a combination of vertical

job segregation and homophily entails that minority employees who are not well represented in management positions may receive lower ratings when manager–employee personal relationships have a positive influence on evaluations and promotions [26, 31, 152].

Stereotype violation has a disparate effect on the "transgressor," depending on their gender. Women are frequently penalized for gender norm violations that make them appear more agentic (e.g., competent and self-confident) and less communal (e.g., kind and warm) than stereotypically expected [115, 195, 247]. This is especially problematic in the hiring domain, where agency is deemed important for leadership positions and it is harder to demonstrate for women without giving an impression of low communion [160].

Workplace proximity and availability of reliable commutes can influence job satisfaction [201] and candidate–employer interactions [262], with amplifying effects introduced by technology [142]. Since discrimination has shaped residential patterns and influenced public transportation, this factor may have disproportionate impacts along racial and ethnic lines [144].

Wage gaps are a key part of power differences between genders and races [40, 81, 121, 269]. This BCF has several concurrent causes, including expectations reinforced by the *status quo* and lower success in salary negotiation [68, 110, 163]. Models that predict the likelihood that candidates will accept an offer can reinforce existing wage gaps.

Social networks topology influences job seekers' awareness of openings [96], and their likelihood of being successfully screened by recruiters due to a successful referral [136, 284]. Well-connected candidates have an inherent advantage over candidates with lower centrality; this advantage may be amplified by algorithms for link prediction [264]. The connection recommender system in a professional social network was found to underperform for women, recommending them less frequently for professional connections than men with similar success rates [39]. This effect may be exacerbated by homophily and biased preferential attachment [14].

3.2 Individual Preferences

Next, we consider BCFs that are an apparent consequence of individual preferences, but represent generalized patterns for protected groups. Listing a bias under this category does not make it an individual responsibility and a reasonable ground for discrimination. On the contrary, we aim to highlight the fact that some seemingly individual choices operated by candidates actually result from wider and recurrent patterns associated with protected attributes. Therefore, employers and providers of algorithmic hiring models should carefully consider these BCFs.

Job satisfaction influences job commitment [20, 228]. Historically disadvantaged groups such as transgender, non-binary, female, black, and disabled workers are more likely to experience discrimination and harassment on the job [225, 228, 261]. This fact may be reflected in datasets as a lower tenure for these groups, which can be penalized by algorithms trained to maximize tenure in order to reduce hiring costs and retain human capital.

Self-promotion gaps related to gender have been documented in the hiring domain [5] and beyond; men tend to self-evaluate higher than women even without intrinsic incentives [87]. This reflects on the visibility and perceived competence of candidates at the different stages of the hiring pipeline; it can be especially difficult to subvert for women due to the social and economic penalties incurred for violating female gender stereotypes [182].

Willingness to commute is another individual factor with gender-related patterns. Women tend to be more restrictive in their choice of job search area [78, 161]. This difference may be related to gender roles with respect to household and childcare responsibilities. Furthermore, people with a migration background are less likely to own a motor vehicle [146].

Salary negotiation differences between men and women are documented, including in propensity [163] and strategy [110]. Different interpretations have been advanced, including lower risk aversion

and the perceived chance of success [110, 122]. Although ostensibly a personal outcome of female candidates, unsuccessful salary negotiation is also based on an unfair *status quo* that influences group expectations [110].

Culture-based avoidance or attraction is a self-selection BCF that influences mainly the sourcing stage. Explicitly mentioning requirements such as community outreach can signal the posture of an employer and act as a pull factor for minority job seekers [192, 234]. On the contrary, job descriptions with unrealistic requirements discourage job seekers who lack one or more requirements, with a repulsive effect that can be stronger for candidates from vulnerable groups [101, 179]. Wording itself can also sustain inequality: gendered wording in job advertisements that makes use of stereotypically agentic (male) language, such as "leadership" or "delivery," can make a position less attractive for women [102]. The same alienating effect can occur by signaling an unwelcoming workplace culture at other stages of the hiring pipeline [274].

Work gaps, i.e., periods without formal employment, often reflect negatively on a candidate's probability of securing a job [101]. Gender asymmetries in caregiving responsibilities put women at a systematic disadvantage [93, 162].

3.3 Technology Blind Spots

Finally, we describe the biases introduced by biased components integrated into larger algorithmic hiring pipelines. This non-exhaustive list aims at demonstrating the need for proactive bias-preventing reasoning also (and perhaps especially) when using off-the-shelf tools.

Ad delivery optimization can skew the audiences reached by job advertisements. Multiple studies have shown that maximization of cost-effectiveness based on ad delivery metrics, such as impressions or clicks, makes delivery skewed in accordance with gender and race stereotypes in jobs [4, 10, 61, 158]. Ad text and images can further skew the audience [4]. This happens even though advertisers design neutral campaigns and is exacerbated if they target specific attributes [258], increasing the opportunity for bias and opacity at the sourcing stage. It should be noted that the platform(s) chosen to run a campaign can introduce a further bias in favor of its predominant demographics; campaigns run on platforms that cater to younger users, for example, are less likely to reach older segments of the population.

Accessibility issues and ableist norms can discourage disabled people from job applications [226] and tilt evaluations against them [245, 246]. AVIs can produce specific patterns from candidates with speech impairment (e.g., short answers) or mismeasure input from candidates with sight impairment (e.g., eye contact), which are judged unfavorably by algorithmic recruitment models [189].

Disparate performance of language processing and computer vision tools has been widely demonstrated with respect to gender, race, and other sensitive attributes [27, 37, 241]. Off-the-shelf algorithms from these domains integrated into hiring pipelines [141] are likely to underserve minority candidates for feature extraction and negatively affect algorithms that are based on these feature.

Differences in platform engagement broadly divide people into frequent and infrequent users. This causes an overrepresentation of the former in the training data, leading to rich-get-richer dynamics in job platforms [185]. Note that minority job seekers may suffer a lower quality of service, leading to disengagement and triggering a negative feedback loop by iteratively lowering training representation, quality of service, and engagement. Moreover, platforms are more likely to have rich profiles and metadata for their common users, while lacking information for less engaged job seekers, who are penalized due to missing data.

Biased psychological assessment performance for minorities can result in systematic disadvantages [212]. Systematic differences in the results of psychometric tests between subgroups [52, 130], caused

by low discriminant validity [133] or construct contamination [8], can be especially problematic for tests embedded in larger resource allocation processes, such as hiring systems.

Background check tools allow employers to obtain additional information on candidates from public records [244]. These databases contain a wealth of pre-conviction information, such as arrest data [157] whose validity as a proxy for crime is questionable [97]. Among other critical aspects, criminal background checks before employment run the risk of feeding the racial disparity of policing [203] and other areas of the criminal justice system [151] into hiring systems.

Interaction biases between employers and hiring technology can amplify tiny differences in algorithmic output [16]. During the sourcing stage, for instance, recruiters tend to focus on very specific credentials, certifications, and keywords that exclude atypical individuals from the initial pool of candidates, despite having the right skills [101]. More general examples include position bias that causes underexposure risks for candidates that are not at the top of a ranking [60] and automation bias leading to over-reliance on technology and reduced human oversight [106, 156].

3.4 Overlaps and Interactions

In the analysis above, we identified and described how the workings of key BCFs contribute to algorithmic discrimination within a digital and non-digital hiring context. It is important to note that these factors are interlocking and overlapping insofar as BCFs are mutually reinforced by the structure of institutions (Section 3.1), individual preferences (Section 3.2), and technological blind spots (Section 3.3).

Intersectional identities are particularly vulnerable as they are likely to be affected by multiple BCFs. Let us consider the case of JS, a woman with children and a migration background looking for a job in the hospitality industry. Her chances of securing suitable employment is hindered from the outset as (1) her connections to the local industries are weak and advertisements do not reach her, keeping her unaware of certain possibilities; (2) she finds out about certain openings but is discouraged from applying by a difficult commute based on public transport. Through public employment services, her data are entered into a shared database accessed by HRs companies. (3) Her profile is hastily filled in and lacks important metadata; this fact pushes her down in rankings every time her profile matches recruiters' queries, also because the time of last login influences job seekers' rankings. (4) Recruiters seek a catering and hospitality degree that JS does not have, making matches with her profile infrequent. (5) Due to a lack of leadership roles in her profile, the only matches she receives are at the apprentice level and (6) she is screened out from openings employing AVI models that do not recognize her accent. She is interviewed by few businesses; most turn her down because (7) they deem her a poor cultural fit and (8) are afraid that she will apply for parental leave. One business is finally willing to hire her through an HR company. (9) Despite a mid-level payment agreement between the employer and the HR company, the latter estimates that JS will accept the minimum wage (which she does) and pockets the difference.

This fictitious, yet plausible story sketches how multiple biases compound and reinforce each other. However, even if intersectional discrimination occurs, there are still reasons why the right approach to technological design can reduce bias. One upshot of understanding bias as an inherently intersectional process is that it also offers a way to reduce discrimination. Since the factors that create bias are interrelated and mutually reinforcing, by halting or ameliorating one BCF, we may introduce positive feedback loops on other BCFs. By removing the discriminatory effect of any one factor, we can hope to reduce its influence on the other factors that reinforce each other in a discriminatory way. For example, an increased representation of a minority group in a company can improve job satisfaction, increase their importance and visibility, and encourage other minority applicants in future hiring rounds. With time, this contributes to reducing vertical segregation,

Table 1. Main Notational Conventions Used in This Work

$x \in \mathcal{X}$	A vector of non-sensitive attributes
$s \in \mathcal{S}$	A vector of sensitive attributes
$s = g$	A sensitive attribute value
$s = g^{\mathsf{c}}$	Complement of a sensitive attribute value
$y \in \mathcal{Y}$	Target variable from domain $\mathcal{Y}$
$i = (x_i, s_i, y_i)$	A data point or item
$\hat{y} = f(x)$	A classifier $f : \mathcal{X} \to \mathcal{Y}$ issuing predictions in $\mathcal{Y}$ for data points in $\mathcal{X}$
$f_{\text{soft}}(x)$	A soft classifier, from which predictions $\hat{y} = f(x)$ can be derived through thresholding
$\tau = \text{argsort}(f_{\text{soft}}(x))$	Ranking of items, typically sorted by target estimates
$i = \tau(k)$	Item at rank k in τ
N	Number of items
N_g	Number of items in g
N_g^k	Number of items in g in the top k positions of a ranking
D_g	Desired representation for group g
$h(x)$	A classifier $h : \mathcal{X} \to \mathcal{S}$ issuing predictions in $\mathcal{S}$ for data points
$h_{\text{soft}}(x)$	A soft classifier for sensitive attributes

with minority employees in key positions shaping company culture, and attracting more candidates with a similar background. This optimistic view reflects positive spillover effects [155, 174].

4 Measures

Fairness measures for algorithmic hiring consider different types of systems, including AVI scoring, CV ranking, and advertising algorithms, operating at different stages and on different data modalities. In this section, we present them in a unified notation (Table 1) and discuss their key dimensions, summarized in Table 2. It is worth noting that the next three sections are based on a systematic review of the literature summarized in Appendix A.

4.1 Dimensions of Fairness Measures

We begin with a summary of dimensions that are important in the algorithmic fairness literature; some are established in the literature, such as *flavor* [178] and *conditionality* [275], others are highlighted as important in the hiring domain, such as *granularity* and *interpretability*. It should be noted that we focus on group fairness; we found a single article treating individual fairness [172].

Flavor. Fairness measures can target different properties of an algorithmic decision-making system. The main notions of fairness in hiring are the following.

— *Outcome fairness* [114] is the most common. It looks at predictions from the perspective of candidates, measuring differences in their preferred outcome, such as being screened in, typically corresponding to a positive prediction $\hat{y}$. Systems are deemed fair from an outcome perspective if quantities such as acceptance rates ($\Pr(\hat{y} = 1)$) or true-positive rates ($\Pr(\hat{y} = 1|y = 1)$) are similar between groups.
— *Accuracy fairness* [24] takes a closer perspective on the decision maker, requesting equalization of accuracy-related properties between groups, such as the average absolute error. Measures in this family lack the notion of preferable outcomes and candidate benefits.

Table 2. Main Measures of Fair Hiring and Their Properties

	Used in	Flavor	Cond.	Multi.	Gran.	Norm.	Interp.
skew@k	[103]	Outcome		No	No	✓	D
NDKL	[11, 103]	Outcome		✓	✓	✓	No
DI	[33, 38, 69, 148, 271]	Outcome		✓	No	✓	✓
DD	[222]	Outcome		No	No	✓	✓
RPP	[4, 132, 199]	Outcome		No	No	No	✓
DRD	[283]	Outcome		No	✓	✓	✓
SKL	[199]	Outcome		No	✓	✓	No
LRR	[48]	Outcome	x	No	✓	No	D
TPRD	[63, 119]	Outcome	y	No	No	✓	✓
FNRR	[148]	Outcome	y	No	No	✓	✓
MED	[231]	Outcome	y	✓	✓	✓	✓
RMS	[119]	Outcome	y	✓	No	No	No
xEO	[185]	Outcome	y	No	✓	No	No
MAE	[231, 276]	Accuracy	y	No	✓	✓	No
BCRD	[148]	Accuracy	y	No	No	✓	No
MID	[148]	Accuracy	y	No	No	✓	No
SD	[222]	Impact		No	No	✓	✓
GBS	[128]	Representational		No	n.a	No	✓
sAUC	[33, 119, 120, 194, 196, 222]	Process		No	n.a.	No	n.a.
GTR	[196]	Process	x	No	n.a.	No	n.a.
MIA	[276]	Process	y	✓	No	✓	n.a.

Columns report the fairness dimensions described in Section 4.1.
In the last column, "D" denotes direction-interpretability, for measures which clearly convey whether one group is at an advantage; ✓indicates both direction- and magnitude-interpretability, the latter being assigned to measures intuitively quantifying the advantage; "No" indicates none of the above; "n.a." in process and representational fairness stand for *not applicable*.

—*Impact fairness* relates algorithmic outcomes to downstream benefits or harms for individuals. These measures are rare in the literature, as they require a broader understanding and modeling of the sociotechnical system around the decision-making algorithm.
—*Process fairness* [111] is a notion that considers the equity of the procedure leading to a decision. Related to *procedural justice* [235], process fairness has been operationalized for algorithmic decision-making based on people's approval for the use of a given feature in a specific scenario. In hiring, this has been associated with the predictability of sensitive attributes from non-sensitive ones [33], implying that the absence of information on sensitive attributes in the data will lead to fair decision-making.
—*Representational fairness* [1] relates to stereotyping and biases in representations. In the context of algorithmic hiring, it is especially relevant for the wording of job descriptions which can skew the probability of application for different demographics [102].

Conditionality. Conditional fairness measures accept group differences, as long as they can be attributed to a set of variables deemed acceptable grounds for differentiation. This dimension is

strongly connected to world views [100, 123], and the extent to which differences in variables between sensitive groups are influenced by measurement errors and unjust social structures.

Granularity. The same model can lead to different results if used in different ways, including the use of different thresholds for classifiers or different cutoffs for rankers. Measures with fine granularity take this fact into account by measuring the fairness of a system across different operating conditions, as opposed to coarse-granularity measures which consider a single point of operation.

Normativity. Measures with explicit normative reasoning set a precise target for groupwise quantities, either in absolute terms or relative to each other. Implicit normative reasoning, on the other hand, is typical of measures introduced without a complete discussion of desiderata. Strong normative reasoning is, in general, preferable, but it is not a guarantee on its own. For example, while a measure providing a contextualized and accurate operationalization of a construct defined in the law displays normativity, the measure will inherit all the limitations of the underlying law.

Interpretability. We consider the interpretability of group fairness along two dimensions. Direction-interpretability allows us to immediately understand whether a group is at an advantage or disadvantage by comparing the measure against a threshold. Magnitude-interpretability lets us quantify the (dis)advantage by evaluating the distance from a threshold.

Native Multinarity. A measure is multinary if it accounts for more than two sensitive attribute values. Binary measures can sometimes be extended to the multinary case, whereas truly multinary measures natively account for this occurrence.

4.2 Notation

Let $x \in \mathcal{X}$ denote a vector containing non-sensitive features and let $y \in \mathcal{Y}$ be an unknown target variable, inferred at test time as $\hat{y} = f(x)$. Furthermore, let $f_{\text{soft}}(x)$ denote a scoring function supporting estimates $\hat{y}$ via thresholding and item ranking $\tau = \text{argsort}(f_{\text{soft}}(x))$ via sorting. Furthermore, let $s \in \mathcal{S}$ indicate a vector of sensitive attributes, so that $s = g$ defines a specific protected group, and $s = g^{\text{c}}$ represents its complement. Overall, a data point or item i is indicated as $i = (x_i, s_i, y_i)$ and $i = \tau(k)$ denotes the item at position k in ranking τ. For cardinality, let N denote the total number of data points in a sample of interest, let N_g indicate the number of items in group g, and let N_g^k be the number of items in g among the top k of a ranking τ. Additionally, let $D_g \in [0, 1]$ denote the desired representation for items in g among the ones receiving a favorable algorithmic outcome, e.g., the top-ranked items or the positively classified ones. Finally, let $h : \mathcal{X} \to \mathcal{S}$ denote a classifier issuing predictions in $\mathcal{S}$ and let $h_{\text{soft}}(x)$ indicate its soft scoring version (e.g., issuing posterior membership probabilities).

4.3 Measures

4.3.1 Outcome Fairness. These measures summarize the perspective of candidates by focusing on their preferred outcome; they are the most common.

Skew at k (skew@k) [103] evaluates a ranking at a specific rank k. It computes the logarithm of the ratio between the desired representation of a sensitive group D_g and the actual representation (N_g^k/k).

$$\text{skew@k}_g = \log\left(\frac{N_g^k/k}{D_g}\right). \tag{1}$$

This measure exhibits strong normativity, as it requires a precise definition of target representation D_g for each sensitive group. It is direction-interpretable, as values above zero indicate an advantage for group g, while the presence of the log function hinders its magnitude-interpretability. As a

summary for multiple groups, Geyik et al. [103] consider the maximum and minimum skew values, e.g., max-skew@k = $\max_{g\in\mathcal{S}}$ skew@k$_g$.

Normalized Discounted Cumulative Kullback-Leibler Divergence (NDKL) [11, 103] computes a divergence between the desired distribution of the representation ($D = \{D_g, \forall g \in \mathcal{S}\}$) and the actual one at rank k ($N^k = \{N^k_g/k, \forall g \in \mathcal{S}\}$). The measure consists of KL divergences, calculated at each rank, and aggregated with a logarithmic discount, making it granular. Notice that KL divergence is another way to summarize skew@k for multiple groups, weighted with a logarithmic discount. It is defined as

$$\text{NDKL} = \frac{1}{Z}\sum_{k=1}^{K}\frac{1}{\log_2(k+1)}d_{\text{KL}}(D, N^k), \tag{2}$$

where Z is a normalization constant. Note that several factors contribute to making NDKL difficult to interpret. First, advantages and disadvantages for a group g at some rank k can yield the same KL divergence value. Second, advantages and disadvantages at different ranks do not compensate each other; if a group is under-represented at low ranks, its overrepresentation at high ranks leads to even worse values of NDKL, despite being a form of compensation.

Acceptance Rate Ratio, better known as **Disparate Impact (DI)** [33, 38, 120, 148], is a measure of disparity in classification with a fixed threshold. Note that this measure is also used in rankings by translating a threshold (or percentile) into a cutoff rank k. This measure is related to US labor law, adverse impact, and the 80% rule (Section 8.2). Below, we report its popular min-over-max version, computing the selection rate ratio between the worst-off and best-off groups:

$$\begin{aligned}\text{DI} &= \frac{\min_{g\in\mathcal{S}} N^k_g/N_g}{\max_{g\in\mathcal{S}} N^k_g/N_g}\\ &= \frac{\min_{g\in\mathcal{S}} \Pr(\hat{y}=1|s=g)}{\max_{g\in\mathcal{S}} \Pr(\hat{y}=1|s=g)}.\end{aligned} \tag{3}$$

We consider this measure interpretable, so long as the best- and worst-off groups are reported.

Acceptance Rate Difference better known as **Demographic Disparity (DD)** [222] focuses on disparities in groupwise acceptance rates, similarly to DI, but computes the difference instead of the ratio. We define it below for classifiers and ranking algorithms, implicitly assuming a cutoff point at rank k for the latter:

$$\begin{aligned}\text{DD} &= N^k_g/N_g - N^k_{g^c}/N_{g^c}\\ &= \Pr(\hat{y}=1|s=g) - \Pr(\hat{y}=1|s=g^c).\end{aligned} \tag{4}$$

This approach may be easier to interpret [286], but fails to capture large disparities when acceptance rates are small. If, for example, $\Pr(\hat{y}=1|s=g) = 10\text{e}{-02}$ and $\Pr(\hat{y}=1|s=g^c) = 10\text{e}{-07}$, we measure a low value DD $\simeq 10\text{e}{-02}$, despite a difference of five orders of magnitude in acceptance rates.

Representation in Positive Predicted (RPP) [4, 132, 199] is a measure used in non-cooperative audits, where the overall population of interest for an algorithm is unknown. In online advertising, campaigners are informed about the size and composition of the audience reached by an ad (positive predicted class) but have no information on platform users and users who were active and thus candidates for an ad impression. This makes it impossible to compute groupwise acceptance rates ($\Pr(\hat{y}=1|s=g)$), but allows for estimates of groupwise representation in the positive predicted

class ($\Pr(s = g|\hat{y} = 1)$) and, trivially, its difference with respect to the complementary group:

$$\text{RPP}_g = \Pr(s = g|\hat{y} = 1) \tag{5}$$

$$\text{RPPD} = \text{RPP}_g - \text{RPP}_{g^c} = 2\,\text{RPP}_g - 1. \tag{6}$$

True-Positive Rate Difference (TPRD) [63, 119] measures disparities in true-positive rates (also known as *recall*) between different sensitive groups. This measure is closely related to equal opportunity [114] and separation [17] and presupposes the availability of a ground-truth variable y to condition on:

$$\begin{aligned} \text{TPR}_g &= \Pr(\hat{y} = 1|y = 1, s = g) \\ \text{TPRD} &= \text{TPR}_g - \text{TPR}_{g^c}. \end{aligned} \tag{7}$$

A closely related measure is the **False-Negative Rate Ratio (FNRR)** [148], defined as

$$\begin{aligned} \text{FNR}_g &= \Pr(\hat{y} = 0|y = 1, s = g) \\ \text{FNRR} &= \frac{\text{FNR}_g}{\text{FNR}_{g^c}}. \end{aligned} \tag{8}$$

As an aggregate measure for non-binary sensitive attributes, Hemamou and Coleman [119] propose the RMS of the vector containing all TPRD components:

$$\text{RMS} = \sqrt{\frac{1}{|\mathcal{S}|}\sum_{g\in\mathcal{S}} \text{TPRD}_g^2}. \tag{9}$$

Nandy et al. [185] propose an **eXtended Equality of Opportunity (xEO)** measure, which is a granular version of TPRD. They consider a soft classifier with a variable threshold and compute the Kolmogorov-Smirnov distance between the resulting score distributions for positives in different sensitive groups.

$$\text{xEO} = \max_x [|\Pr(f_{\text{soft}}(x) \leq t|y = 1, s = g) - \Pr(f_{\text{soft}}(x) \leq t|y = 1, s = g^c)|]. \tag{10}$$

Discounted Representation Difference (DRD) [283] is a granular measure of DD focused on rankings. It measures the difference in acceptance rates with a variable cutoff rank k by observing the sensitive attribute of the item $\tau(k)$ and applying a logarithmic rank-based discount before updating a counter. In other words, DRD measures groupwise representation as

$$\text{DRD} = \sum_{k=1}^{K} \frac{1}{\log_2(k+1)} [\mathbb{1}(s_{\tau(k)} = g) - \mathbb{1}(s_{\tau(k)} = g^c)]. \tag{11}$$

This measure is interpretable since it conveys the difference in candidates' exposure across groups under a recruiter browsing model with logarithmic decay. See Carterette [41] for an introduction to browsing models.

Score KL Divergence (SKL) [199] considers the distribution of scores ($f_{\text{soft}}(x_i)$) in different groups and measures unfairness by computing their KL divergence. Let D_g^f define the probability distribution of continuous scores $f_{\text{soft}}(x_i)$ for group g, then SKL is defined as

$$\text{SKL} = \text{KL}(D_g^f, D_{g^c}^f). \tag{12}$$

Log Rank Regression (LRR) [48] is a measure defined for search engines by fitting a linear model on the logarithm of rank $k = \tau^{-1}(i)$ at which candidates are presented on the result page. Independent variables comprise both sensitive s and non-sensitive attributes x, where the latter include skills and education. LRR reports the coefficient (and p-value) associated with the sensitive attribute. Fitting models to quantify the influence of a sensitive feature on an outcome variable

is typical of situations with limited access to the model(s) responsible for the outcomes. This is a common setting for external non-cooperative audits.

$$\log(\tau^{-1}(i)) = \beta_x x_i + \beta_s s_i + \mu + \epsilon$$
$$\text{LRR} = \hat{\beta}_s. \tag{13}$$

A similar approach is proposed in Lambrecht and Tucker [158] to estimate the influence of sensitive attributes such as age and gender in advertisement delivery optimization. This measure considers differences in outcomes acceptable, so long as they can be explained by non-sensitive attributes x.

Mean Error Difference (MED) [231] focuses on regression, measuring systematic groupwise biases. Following a more general paradigm, MED caters to the multinary case by considering the maximum and minimum (signed) bias.

$$\text{MED} = \max_{g \in \mathcal{S}} \frac{1}{N_g} \sum_{i \in g} (y_i - f_{\text{soft}}(x_i)) - \min_{g \in \mathcal{S}} \frac{1}{N_g} \sum_{i \in g} (y_i - f_{\text{soft}}(x_i)). \tag{14}$$

4.3.2 Accuracy Fairness. These measures study model accuracy across sensitive groups, aligning more closely with the perspective of decision makers. They are all conditional on the target variable y, inheriting the biases encoded in the ground truth.

Mean Absolute Error (MAE) [231, 276] targets regression problems similarly to MED (Equation (14)). It compares groupwise accuracy by measuring the absolute error for each individual and computing the average for items in the same sensitive group.

$$\text{MAE} = \frac{1}{N_g} \sum_{i \in g} |y_i - f_{\text{soft}}(x_i)| - \frac{1}{N_{g^c}} \sum_{i \in g^c} |y_i - f_{\text{soft}}(x_i)|. \tag{15}$$

Balanced Classification Rate Difference (BCRD) [148] is a measure of the disparity in classification accuracy between groups. It targets the balanced classification rate, defined as the average between the true-positive and the true-negative rate:

$$\text{BCR}_g = \frac{\text{TPR}_g + \text{TNR}_g}{2}$$
$$\text{BCRD} = \text{BCR}_g - \text{BCR}_{g^c}. \tag{16}$$

Mutual Information Difference (MID) [148] is another measure of classification accuracy fairness. It computes model accuracy within a group g as the **Mutual Information (MI)** between the target y and the prediction $\hat{y}$ for the items in g and compares it against the MI for the remaining items.

$$\text{MI}_g = \text{MI}_{\{i \in g\}}(\hat{y}_i, y_i)$$
$$\text{MID} = \text{MI}_g - \text{MI}_{g^c}. \tag{17}$$

4.3.3 Impact Fairness. This flavor of fairness models the impact of algorithmic outcomes on data subjects, measuring differences in harms and benefits between populations.

Salary Difference (SD) [222] is the only measure of this family proposed in the literature. Considering an embedding-based model, each candidate is matched to the closest vacancy in the embedding space. Based on the average wage for the role described in the vacancy, an average salary is calculated for male and female candidates in the same industry, and their difference quantifies

the gender impact of the model on earnings in that industry.

$$f(x_i) : \text{closest job match for candidate } i$$
$$W(f(x_i)) : \text{average wage for job} f(x_i)$$
$$\text{SD} = \frac{1}{N_g} \sum_{i \in g} W(f(x_i)) - \frac{1}{N_{g^c}} \sum_{i \in g^c} W(f(x_i)). \tag{18}$$

4.3.4 Representational Fairness. Measures to quantify stereotypes and representational harms are relatively understudied [1, 92], despite being a fundamental driver in the reinforcement of societal biases.

Gender Bias Score (GBS) [128] is a coarse measure of bias toward "masculine" or "feminine language" in job descriptions. It takes advantage of the word inventory from Konnikov et al. [150], which comprises traits associated with gender roles in labor, so that each term t in a job description can theoretically be coded as belonging to a set of stereotypically masculine, feminine, or neutral traits. The measure is defined as

$$\text{GBS} = \text{sign}(x_m - x_f) \cdot \max \left\{ \frac{x_m - x_f}{x_m}, \frac{x_f - x_m}{x_f} \right\}, \tag{19}$$

where x_m (x_f) represents the number of stereotypically male (female) words in a job description.

4.3.5 Process Fairness. These notions of fairness operationalize desiderata about decision-making beyond outcomes. Frequently, they derive from judgments of admissibility for specific variables x in a given scenario. In algorithmic hiring, the literature focuses on the amount of information on sensitive attributes s encoded by non-sensitive features x and the target variable y, with the goal of minimizing it.

Sensitive Area Under the Receiver Operating Characteristic Curve (sAUC) [33, 119, 120, 194, 196, 222] is a widely used measure of the information about sensitive attributes stored in (proxy) non-sensitive attributes. It is based on training a classifier $h(x)$ for the sensitive attribute s on non-sensitive features x; its accuracy is evaluated as the AUC. If little information on sensitive attributes can be recovered from the variables employed in the decisions, then the decision is deemed procedurally fair according to this definition.

$$\text{sAUC} = \text{AUC}(h_{\text{soft}}(x)). \tag{20}$$

Hemamou et al. [120] extend (sAUC) to the multi-class case by training one-vs-all classifiers and reporting their maximum AUC value.

Ground Truth Regression (GTR) [196] regresses (x, s) on y, and focuses on the latter coefficient ($\hat{\beta}_s$) similarly to LRR (Equation (13)), with the key difference that the target of the regression is the target variable. This measure estimates the importance of the sensitive attribute to predict the target, conditional on non-sensitive attributes. In other words, GTR audits the biases encoded in the target variable, recognizing it as a key factor for the decision-making process, regardless of the predictive model employed:

$$\log \left(\frac{y_i}{1 - y_i} \right) = \beta_x x_i + \beta_s s_i + \mu + \epsilon$$
$$\text{GTR} = \beta_s. \tag{21}$$

Both sAUC and GTR are *ex-ante* measures, i.e., they can be computed from the data before the final model is trained; therefore, the *granularity* dimension does not apply to them.

Mutual Information Amplification (MIA) [276] is another measure of process fairness, which considers the process fair if it does not reveal more information about sensitive attributes than the

ground truth does. This metric computes the MI between first $\hat{y}$ and s, and second between y and s; positive differences between the two are considered an undesirable amplification of information about sensitive attributes leaked through predictions.

$$\text{MIA} = \text{MI}(\hat{y}, s) - \text{MI}(y, s). \tag{22}$$

Differently from sAUC and GTR, MIA is an *ex-post*, model-dependent measure.

4.4 Discussion

In this section, we describe the main characteristics of the surveyed measures and how they inform fairness measurement choices.

Choosing a Measure. We introduce a made-up scenario to exemplify deliberation around fairness monitoring. *EquiHire*, a fictitious HR company, is choosing a fairness measure to monitor its candidate search systems for fairness with respect to ethnicity. For the sake of simplicity, they restrict the realm of possibilities to Table 2. Since ethnicity is a multinary sensitive attribute, they discard binary measures such as skew@k and DD. Moreover, EquiHire practitioners want to describe their equity strategy in white papers and external communication. Therefore, they prioritize an interpretable measure, excluding complex measures such as NDKL and SKL which would be difficult to explain. Although they have access to signals related to candidate fitness (y), they become available with great delay after hiring decisions are taken. To assess system fairness in a timely manner, they discard y-conditional measures such as TPR and RMS. Finally, since they have visibility into the downstream hiring decisions, they can use a low-granularity measure. These considerations lead EquiHire to choose DI as their primary fairness measure. For a complementary angle focused on outcome fairness, they choose to monitor wage differences for employees hired through their systems, adopting SD as an additional measure.

To reiterate, this is just one fictitious scenario and not a fixed recommendation for choosing the right measure. It demonstrates the importance of key characteristics of the measures summarized in Table 2 for practitioner decisions; in the paragraphs below, we expand on the factors driving these decisions.

Fairness Diagnostics vs. Fairness Optimization. Ease of interpretation is a desirable property, especially for fairness measures, as confirmed by the popularity of DI. Nevertheless, several fairness measures adopted in hiring are difficult to read in absolute terms, making it challenging to understand the severity of fairness violations based on their values. These values can only be interpreted relative to one another, i.e., it is clear that one value is more desirable than another, making them a possible target for optimization but less suited for diagnostics. This is especially true for measures of accuracy fairness such as BCRD and MID. Measures of outcome and impact fairness, such as DD and SD, tend to be more interpretable, since they are typically defined as differences between the probability of obtaining desirable outcomes for different groups. This holds especially for binary measures, while multinary measures, such as RMS, often sacrifice interpretability to summarize disparities across multiple groups.

Blurred Lines between Ranking and Classification. In the literature, classification and ranking are typically considered separate tasks with separate algorithmic fairness measures. In hiring, we find ranking systems evaluated according to classification measures such as DI [38], selection tasks solved with rankers [38], and target variables of similar nature encoded as binary, multinary, or continuous (Section 6). For this reason, the separation between classification and ranking is less rigid than in other domains, and the question of which measures are best suited to evaluate these systems arises. A fundamental part of the answer lies in the expected flexibility of use for these models. If the key aspects of their usage are fixed, including the cutoff thresholds for different outcomes, then a measure with coarse granularity, considering a single operating condition (typical of the

fair classification literature—e.g., DD, TPRD), is most suited for evaluation. On the contrary, for models whose operating conditions are more uncertain, requiring human interaction or threshold fine-tuning, it is preferable to adopt a measure with finer granularity (often derived from the fair ranking literature—e.g., NDKL, DRD), which averages outcomes across multiple realizations of a browsing model.

Diagnosing Biases. Fairness measures are especially useful when they provide actionable diagnostics by highlighting specific biases described in Section 3. Measures of process fairness such as sAUC and MIA detect the presence of sensitive attribute proxies. Dedicated proxy reduction approaches (Section 5) can mitigate the resulting risk of discrimination. Representational fairness (GBS) captures biases in representations; when applied to job descriptions, extreme values highlight stereotypically gendered language, signaling a risk of culture-based avoidance. Impact fairness can highlight nuanced aspects of hiring beyond selection rates; SD, for instance, captures wage gaps and salary negotiation differences prompting further analysis into remuneration packages. Measures of outcome (e.g., DI) and accuracy fairness (e.g., MAE) can serve as high-level metrics by highlighting the effects of multiple factors acting together. To exemplify, a combination of job segregation, work gaps, and disparate performance of language processing tools, even if relatively weak on their own, may trigger a very low (unfair) value of DI when they compound.

Assorted Fairness Flavors. No single measure provides a complete picture. Different fairness flavors have complementary strengths and weaknesses. Most of the fairness measures considered in hiring focus on outcome equity; the field is strongly influenced by the 80% rule (Section 8.2), which is appreciated as a quantitative rule of thumb, but also often criticized [74, 267]. Indeed, outcome fairness is based on a narrow view of algorithmic hiring as a single decision point abstracted away from its context, where equity is purely a function of algorithmic estimates and (sometimes) their similarity to target variables approximating a ground truth. Similar criticism can be moved to accuracy fairness, with the additional limitation that it is less interpretable and less aligned with the desiderata of job seekers. Departing from this approach, process fairness is operationalized with reference to the information on sensitive attributes encoded in the remaining variables. More work is required to understand process fairness in hiring more broadly and to suitably (if at all) operationalize it quantitatively. Impact fairness quantifies the downstream impacts of algorithmic outcomes on different populations; modeling important aspects of the surrounding socio-technical system, it goes beyond brittle algorithmic abstraction. Since these measures are exceedingly rare, we highlight the need for more context-specific research to understand and model the benefits and harms of decisions on job candidates at different stages of the hiring pipeline. Overall, practitioners developing equity monitoring protocols should consider multiple fairness angles to gain a nuanced and more complete picture.

Ignoring Privileged and Disadvantaged Groups. The definition of sensitive attributes in anti-discrimination law is informed by the historical (dis)advantage of specific groups [230]. Gender and race, for example, are considered sensitive attributes due to the recurrent structural disadvantages incurred by women and black people. This is especially true in hiring, where surveys and meta-analyses find consistent biases against women and ethnic minorities [23, 147, 208]. Although it is commonly held, especially in fairness research, that disparities against disadvantaged groups should be mitigated, there is less consensus on how to treat algorithms that happen to favor historically disadvantaged communities. In other words, should algorithmic fairness be symmetrical and reject *a priori* notions of privileged and disadvantaged groups? This is an important normative question, seldom acknowledged in the fairness literature, with immediate consequences on the choice of measures. For example, the version of DI reported in Equation (3) computes the ratio between the acceptance rates of the worse-off group (min) and the best-off group (max), ignoring prior knowledge of structural inequality and affected communities.

Table 3. Bias Mitigation Methods to Improve Fairness in Hiring

	In	Family	Measures	s availability	Mode	Approach	Summary
Rule-based Scraping or Substitution	[63, 194, 222]	PRE	sAUC, DD, TPRD, SD	None	Text	Proxy reduction	Remove or substitute gender identifiers based on hard-coded rules
Importance-Based Scraping	[33, 194]	PRE	sAUC, DI	Train		Proxy reduction	Iteratively remove proxy features most predictive of sensitive attribute
Balanced Sampling	[11, 276]	PRE	NDKL, MIA, MAE	Train		Re-balancing	Resample training set with same group cardinality in each class
Group Norming	[33]	PRE	sAUC, DI	Train		Groupwise feature transform	z-score normalization of each feature within each group
Subspace Projection	[194]	PRE	sAUC	None	Text	Proxy reduction	Remove gender information via embedding projection
Adversarial Inference	[120, 199, 222, 276]	IN	sAUC, DD, TPRD, SD, MIA, MAE, SKL	Train		Proxy reduction	Reduce sensitive info in latent representation through additional adv. loss
Face Decorrelation	[120]	IN	sAUC, DI	None	Image	Proxy reduction	Discourage intermediate representations correlated with face ID through adv. loss
Name Decorrelation	[119]	IN	sAUC, TPRD, RMS	None	Text	Proxy reduction	Discourage intermediate representation correlated with names by minimizing MI
Fair TF-IDF	[69]	IN	DI	Train	Text	Proxy reduction	Reduce feature weight according to its group-specificity
DetGreedy	[11, 103, 240]	POST	NDKL	Run-time		Output re-ranking	Re-rank items enforcing desired group representation in each prefix
CDF Rescoring	[185]	POST	xEO	Run-time		Output re-rescoring	Re-score items with groupwise CDF trick
Spatial Partitioning	[38]	POST	DI	Train		Output re-ranking	Promote candidates based on group membership probability

CDF, Cumulative Distribution Function.

5 Mitigation Strategies

Several algorithms have been proposed in the literature to improve model fairness; they are summarized in Table 3. We present them distinguishing between pre-processing, in-processing, and post-processing algorithms, depending on their applicability before, during, and after training.

5.1 Pre-Processing

Rule-based approaches perform a set of manipulations, typically defined by experts, on text data. They are related to process fairness (Section 4), as they focus on reducing the amount of information on sensitive attributes contained in non-sensitive features, i.e., to remove sensitive attribute proxies (Section 3). *Rule-based scraping* [63] is a heuristic for *text data* focused on gender, aimed at removing all words that explicitly refer to the gender of a person, including first names and titles. In general, it acts as a feature transformation $x' = \text{scrp}(x)$. Under bag-of-words representations, the scraping function is a feature selection mechanism that removes features in a censored vocabulary $\mathcal{V}$.

$$\text{scrp} : \mathcal{X} \to \mathcal{X}' \subseteq \mathcal{X}$$
$$\text{scrp}(x^j) = \begin{cases} x^j, & \text{if } x^j \notin \mathcal{V} \\ \emptyset, & \text{if } x^j \in \mathcal{V} \end{cases}.$$

De-Arteaga et al. [63] study the effectiveness of this approach in *job classification* from short biographies. They find it to be moderately effective in reducing TPRD. Parasurama and Sedoc [194]

take this approach one step further on a CV screening application. They define several rules for removing other strings related to gender, including e-mail addresses, intrinsically gendered words (e.g., "waitress"), and hobbies. *Rule-based substitution* [222] is a closely related approach, which replaces intrinsically gendered words with neutral words; for example, "his" is changed into "theirs." Due to redundant encoding of sensitive information in the data, rule-based approaches often represent a weak baseline with limited impact on fairness.

Importance-based scraping is an adversarial approach for feature removal based on a sensitive attribute classifier. Parasurama and Sedoc [194] consider a *resume screening* system and exploit contextualized word representations to train a gender classifier and iteratively scrape the words with the largest feature importance for gender classification. In other words, they train a sensitive attribute classifier

$$\hat{s} = h(x)$$

and assess the importance of features by selectively removing them from the prediction task. Proxy features $\mathcal{P}$ found to be most predictive of s (highest marginal contribution to $h(\cdot)$, e.g., measured via SHAP) are scraped:

$$\text{scrp}(x^j) = \begin{cases} x^j, & \text{if } x^j \notin \mathcal{P} \\ \emptyset, & \text{if } x^j \in \mathcal{P} \end{cases}$$

Parasurama and Sedoc [194] show that SHAP-based scraping can achieve a sizeable sAUC drop with a limited negative impact on performance. Booth et al. [33] develop an equivalent scheme for iterative feature removal in AVI analysis from multimodal data. They confirm the suitability of this method to improve process fairness (sAUC); outcome fairness (DI) also improves, but only for systems that were initially very unfair.

Subspace projection is a common approach to reduce gender bias in *text* representations based on word embeddings. Initially proposed by Bolukbasi et al. [32], this method is based on the observation that most intrinsically gendered information in word embeddings, such as the difference between the vectors for "mother" and "father," is contained within a small gender subspace. The algorithm is based on re-embedding each word by projecting it orthogonally to this space. Let $\vec{w}$ denote the embedding of a word and G the gender subspace. Each word is re-embedded to

$$\vec{w}' = \frac{\vec{w} - \text{proj}_G \vec{w}}{||\vec{w} - \text{proj}_G \vec{w}||}.$$

In the algorithmic hiring literature, subspace projection is used in Parasurama and Sedoc [194] to reduce gender biases in a *resume screening* application. Although in theory this approach is suitable to remove gender proxies and contrast the negative effects of stereotype violations, it proves ineffective in reducing the amount of gender information contained in resumes, as measured by sAUC, in alignment with prior art [108].

Balanced sampling is a broadly applicable scheme to reduce correlations between sensitive attributes and target variables in the training set. Arafan et al. [11] propose a down-sampling scheme to achieve an equal representation of sensitive groups among positive and negative points in the training set. For a binary sensitive attribute, this condition can be formulated as

$$\Pr_{\sigma}(s_i = g | y_i = \overline{y}) = \Pr_{\sigma}(s_i = g^c | y_i = \overline{y}), \forall \overline{y} \in \mathcal{Y},$$

where we let σ denote an algorithm's training set. Yan et al. [276] propose an upsampling approach to enforce the same condition before training a multimodal data fusion algorithm for the analysis of AVIs [141]. Overall, balanced sampling can reduce the influence of job segregation on hiring algorithms.

Group norming is a feature manipulation approach for *tabular data* enforcing a similar feature distribution in all sensitive groups, through normalization. Booth et al. [33] propose *Groupwise z-Normalization*, i.e., they divide data points based on their sensitive group membership, and standardize each feature by subtracting the groupwise mean and dividing by the groupwise standard deviation.

$$\mu_g = \frac{1}{N_g} \sum_{i \in g} x_i$$
$$\text{std}_g = \frac{1}{N_g} \sum_{i \in g} (x_i - \mu_g)^2$$
$$x'_i = \frac{x_i - \mu_g}{\text{std}_g} \text{ if } i \in g.$$

This approach represents an intermediate view on conditional discrimination: it prohibits *inter-group* discrimination based on a specific feature, while allowing it as a basis for *intra-group* discrimination. It is tested on multimodal data and found to decrease gender predictability (sAUC) for paraverbal and visual data, while increasing predictability for verbal features and only marginally improving DI. It is unclear how to generalize this method under multiple protected attributes (e.g., race and gender); a separate application to each intersectional group (e.g., black women) is the most straightforward extension, but increasingly smaller groups run the risk of unstable normalization. Considering its limited impact on both process and outcome fairness, in conjunction with its controversy in US employment law [22, 165], the use of group norming is not recommended.

5.2 In-Processing

Adversarial inference [120, 199, 222, 276] is an in-processing approach to remove sensitive information from latent representations that can be applied across tasks (e.g., classification, ranking) and data modalities (e.g., tabular, images). It leverages process fairness to reduce sensitive attribute proxies in pursuit of outcome fairness [73]. This approach directly models an adversary trying to infer an individual's sensitive attributes s_i from their latent representation $l(x_i)$. The latent representation is typically derived in one or more layers of a neural architecture whose goal is to predict the target variable y associated with employability. The adversarial loss for inference is defined as

$$L_{\text{INF}}^{\text{ADV}} = \frac{1}{N} \sum_i \text{dist}(s_i, d(l(x_i))),$$

where $d(\cdot)$ represents the layer(s) in the adversarial branch, so that $h(x) = d(l(x_i)) = \hat{s}_i$ is the inferred sensitive attribute value, and $\text{dist}(\cdot)$ computes its distance from the actual value s_i. Rus et al. [222] show the suitability of this method to improve selected process, outcome, and impact fairness indicators in a *job recommendation* scenario, while Peña et al. [199] demonstrate SKL improvements in a multimodal setting based on synthetic *resume ranking*.

Face decorrelation [120] was proposed for AVI systems and, more generally, algorithms that process *face images*. This approach is based on the key assumption that face data contain enough information on sensitive features such as gender and race, so that successful debiasing can be achieved without explicit sensitive attribute information. More in detail, let $l(x)$ denote a latent representation from a neural architecture used for employability prediction, and let $w(x)$ denote a representation of the candidates' face obtained from a state-of-the-art feature extraction method for face recognition [65]. Hemamou et al. [120] propose two schemes based on MSE and **Negative Sampling (NS)**. Under MSE, the adversary branch tries to leverage latent representations to

reconstruct face features by minimizing

$$L_{\text{MSE}}^{\text{ADV}} = \frac{1}{N}\sum_i [d(l(x_i)) - w(x_i)]^2,$$

where $d(\cdot)$ represents the final dense layer in the adversarial branch. Under NS, the adversary exploits the latent representation $l(x_i)$ extracted from an interview to discriminate the respective candidate from the remaining ones by maximizing the softmax

$$L_{\text{NS}}^{\text{ADV}}(x_i) = \frac{\exp(\text{sim}(l(x_i), w(x_i)))}{\sum_{j\neq i}\exp(\text{sim}(l(x_i), w(x_j)))},$$

where $\text{sim}(\cdot)$ represents a similarity function between the face recognition features $w(x_i)$ and the representations $l(x_i)$ learnt by the main branch of the network. Both variants of adversarial removal (NS, MSE) are shown reduce the sensitive information encoded in latent representations: notably, their sAUC is on par with adversarial methods explicitly trained to predict gender and ethnicity and may be suited for settings where sensitive attributes are unavailable during training. The effectiveness of face decorrelation for outcome fairness is more nuanced, showing positive effects on DI for very unfair models based on the video modality and limited effects on more equitable systems based on language and audio.

Name decorrelation [119] is a similar approach proposed for *text* data, based on sensitive information encoded in *names*. The goal of this method is to reduce the MI between the representations of individuals, such as document embeddings extracted from their resumes, and a word embedding of their name. To handle the complexity of MI estimation in high-dimensional continuous spaces, the method focuses on the MI between the latent representation of the input $l(x_i)$ and a low-dimensional projection of their name $\tilde{t}_i$. The adversarial loss function is defined as

$$L_{\text{name}}^{\text{ADV}} = \frac{1}{N}\sum_i \hat{\text{MI}}(\tilde{t}_i, l(x_i)).$$

This approach reduces the ability of adversaries to infer an individual's gender and ethnicity from their hidden representations as measured by sAUC in a *job classification* task. In terms of outcome fairness, name decorrelation is found to improve TPRD and RMS with respect to gender but not ethnicity. It is worth noting that the original disparities achieved by a vanilla model, measured by RMS and TPRD, were smaller for ethnicity than for gender. Similarly to Hemamou et al. [120], this result suggests that targeting this type of process fairness can improve outcome fairness in highly imbalanced situations, while only providing limited benefits in situations with lower inequity.

Feature weighting schemes learn to decrease the importance of a feature based on its likelihood of increasing the unfairness of a model. *Fair TF-IDF* [69] is an in-processing approach for *text* classification and ranking applications, such as *resume filtering*. As the name suggests, this algorithm is an extension of TF-IDF [237], one of the most popular and influential algorithms in text search engines. In response to a query describing a job, TF-IDF produces a score $f_{\text{soft}}(x_i)$ for each resume based on the count of query terms in item i and on the specificity of the terms in the entire resume collection. In other words, let t denote a term (word), potentially present in a resume i and in a query q describing a job posting; $\text{TF-IDF}(i,t)$ is the product of a term frequency $\text{tf}(i,t)$ and an inverse document frequency $\text{idf}(t)$. The $\text{tf}(i,t)$ factor summarizes the importance of t in resume i by counting the occurrences of t in i. The $\text{idf}(t)$ factor conveys the specificity of term t by counting how many resumes in the entire collection contain the term t and defining $\text{idf}(t)$ as its inverse; $\text{idf}(t)$ can be seen as a weight that increases (decreases) the importance of rare (common) words.

Deshpande et al. [69] propose a further weighting scheme based on penalizing terms that are highly specific to a given group.

$$\begin{aligned} p\text{-ratio}(t) &= \frac{\min_{g \in \mathcal{S}} \Pr(t \in i | i \in g)}{\max_{g \in \mathcal{S}} \Pr(t \in i | i \in g)} \\ \text{fair-tf-idf}(i,t) &= \text{tf(i,t)} \cdot \text{idf(t)} \cdot p\text{-ratio(t)} \\ f_{\text{soft}}(i,q) &= \sum_{t \in q} \text{fair-tf-idf}(i,t). \end{aligned}$$

The authors also propose more complex weighting schemes for the fairness factor p-ratio(t) and test their ability to improve system fairness as measured by DI.

5.3 Post-Processing

DetGreedy [103] and its variants are post-processing methods for *ranking* algorithms targeting NDKL. Given a desired representation of sensitive groups, expressed as a target distribution $D = \{D_g, \forall g \in \mathcal{S}\}$ (Equation (2)), Geyik et al. [103] seek to ensure this representation at different ranks. Starting from the target distribution D, DetGreedy populates the ranking progressively from the top to the bottom rank with the most relevant items from under-represented groups. To do so, it maintains a counter N_g^k for the number of items from group g that have already been placed in the top k positions of the ranking; this counter determines two priority sets from which items at rank $k+1$ can be drawn. The high priority set $\mathcal{G}_H$ consists of items from groups that are below the desired quota. The low-priority set $\mathcal{G}_L$ consists of groups that are above their desired quota, but only by a small margin.

$$\begin{aligned} \mathcal{G}_H &= \{i \in g : N_g^k < \lfloor D_g \cdot (k+1) \rfloor\} \\ \mathcal{G}_L &= \{i \in g : \lfloor D_g \cdot (k+1) \rfloor \leq N_g^k < \lceil D_g \cdot (k+1) \rceil\}. \end{aligned}$$

If $\mathcal{G}_H$ is not empty, DetGreedy chooses the most relevant item from the groups in $\mathcal{G}_H$; otherwise, it samples one from $\mathcal{G}_L$.

$$\tau(k+1) = \begin{cases} \text{argmax}_{i \in \mathcal{G}_H, i \notin \{\tau(1),\ldots,\tau(k)\}} f_{\text{soft}}(x_i), & \text{if } \mathcal{G}_H \setminus \{\tau(1),\ldots,\tau(k)\} \neq \emptyset \\ \text{argmax}_{i \in \mathcal{G}_L, i \notin \{\tau(1),\ldots,\tau(k)\}} f_{\text{soft}}(x_i), & \text{otherwise} \end{cases}$$

By prioritizing items with higher scores in $\mathcal{G}_L$, DetGreedy may end up violating some minimum representation constraints at some rank, i.e., $\exists g$ s.t. $N_g^k < \lfloor D_g \cdot k \rfloor$. This happens when more than one group is in $\mathcal{G}_H$. To mitigate this risk, Geyik et al. [103] propose two variants termed *DetCons* and *DetConsSort*. DetCons tries to avoid this occurrence by prioritizing groups that are more likely to enter $\mathcal{G}_H$ at the next iteration, while DetConsSort is a non-greedy algorithm that can re-order previous items dynamically to avoid constraint violations at the current rank. DetGreedy is implemented in *LinkedIn Recruiter* to ensure equitable gender representation in *candidate search*; the desired gender distribution D_g is made query-dependent and set to match the distribution of qualified candidates for the search criteria.

Cumulative Distribution Function (CDF) rescoring [185] is a post-processing method targeting xEO for *recommender systems*. xEO is a fine-granularity measure defined in Equation (10), studying the properties of the soft classifier $f_{\text{soft}}(x)$ at every possible threshold; it requires that the probability of positive items in a group $\{i \in g : y_i = 1\}$ achieving a score below a threshold t should be the same for every group, at every threshold t. This is achieved by re-mapping item scores to their groupwise CDF as

$$f'_{\text{soft}}(x_i) = \Pr(f_{\text{soft}}(x) \leq f_{\text{soft}}(x_i) | y = 1, s = s_i).$$

In other words, new soft scores are mapped to the $[0, 1]$ interval, according to the CDF of old scores for positive points in their group, ensuring xEO = 0. Nandy et al. [185] propose several extensions to this approach, to achieve xEO across all target class values, to account for position bias in the target, and to tradeoff fairness and accuracy. This method is tested on a *friendship recommendation* engine in a live proprietary system (most likely *LinkedIn*), where sensitive groups are defined based on the level of activity on the platform, mitigating potential differences in platform engagement.

Spatial partitioning [38] is a heuristic to select an optimal group of applicants σ from screening tests, with the additional difficulty that sensitive attribute values are unknown during testing. First, two estimators for the target ($f_{\text{soft}}(x)$) and protected variable ($h_{\text{soft}}(x)$) are developed on a training set. These estimators are applied to the test set, which is ranked customarily as $\tau = \text{argsort}(f_{\text{soft}}(x))$. The top candidates are selected from the ranking and put into σ. Given the systematic groupwise differences encoded in the data, the selected candidates tend to be mostly from the privileged group. Spatial partitioning mitigates this bias by replacing the candidates in σ who have the lowest values of $[f_{\text{soft}}(x) + h_{\text{soft}}(x)]$—or some other linear combination of the two—with the candidates who have the highest values of $[f_{\text{soft}}(x) + h_{\text{soft}}(x)]$ among the ones that were originally not chosen. This approach aims to rebalance the privileged and disadvantaged group in the final set σ while maintaining good accuracy in the selection of high-performance candidates.

5.4 Discussion

In this section, we present the main trends for fairness enhancement in the algorithmic hiring literature and describe important factors guiding these choices (summarized in Table 3). Section 7 will expand on this analysis presenting additional dimensions that guide fairness mitigation in practice.

Mitigating Biases. Some of the proposed approaches are suited to mitigate specific biases described in Section 3. Balanced sampling, for example, can counter the effect of under-representation in training sets caused by factors such as horizontal and vertical job segregation. Proxy reduction methods (e.g., decorrelation, adversarial inference) target sensitive attribute proxies, reducing the overreliance of models on sensitive information. Group norming seems applicable to counter systematic biases encoded in input features against protected groups, such as biased employee evaluations. However, it is certainly preferable to understand and improve the measurement system that provides these biased features rather than simply matching the distribution of input features across protected groups without a clear understanding of the underlying socio-technical system. Finally, output re-ranking can mitigate the joint effect of different BCFs. Its application comes with risks and opportunities described below.

Opportunities and Risks of Post-Processing. Post-processing approaches, such as DetGreedy, are the easiest to integrate into existing systems, as they can be modularly added to algorithmic pipelines [257]. Indeed, most post-processing methods in algorithmic hiring are proposed and deployed at *LinkedIn*, a large company with an established platform powered by interactions between complex data infrastructure and interdependent algorithmic modules [12, 104]. It is worth noting that post-processing explicitly takes into account sensitive attributes to change algorithmic outcomes for job candidates, which may be critical for disparate treatment and affirmative action (US) as well as direct discrimination and positive action (EU) [22, 112]. Additionally, post-processing requires runtime access to the sensitive attributes of all data subjects, as described in the next paragraph. These are likely two key factors explaining why post-processing is less popular and why hybrid approaches combining different types of fairness interventions, e.g., mitigating via pre- *and* post-processing, remain under-explored [11].

Sensitive Data. Different mitigation strategies have different requirements for sensitive attributes s. Post-processing methods, such as DetGreedy and CDF restoring, typically require knowledge

of sensitive attributes during runtime. Conversely, pre-processing and in-processing approaches, such as importance-based scraping and adversarial inference, require access to sensitive attributes only during training, not testing. Furthermore, specific data modalities, such as text and vision, facilitate specialized methods like subspace projection and face decorrelation, which function without sensitive attribute information about data subjects. Each of these methods has distinct data requirements, ranging from highly restrictive to more lenient. Having access to sensitive attributes at runtime enables precise and reliable interventions, whereas strategies that do not rely on this data must be continuously monitored to ensure they are meeting their intended goals.

Outcome Fairness through Proxy Reduction? A very clear trend in the literature is the popularity of proxy reduction methods, such as adversarial inference, that target process fairness (sAUC) in pursuit of outcome fairness (e.g., DI, TPRD). This approach probably gained popularity because it aligns with legislation against *disparate treatment* (US) and *direct discrimination* (EU), which prohibit basing hiring decisions on protected attributes [2, 158] and because it does not require runtime access to sensitive attributes. This approach is also adopted by vendors, such as *Pymetrics* and *Hirevue*, upon detecting violations of the 80% rule in pre-deployment audits [124, 209]. While this method can mitigate large outcome disparities, it does not produce convincing improvements when the inequity is less significant [33, 119, 120]. Moreover, it is worth noting that, in contrast to post-processing, this approach achieves mitigation on a held-out test set during development but provides no guarantee at deployment. More general studies are required to understand the effects of proxy removal on outcome fairness and ensure actual benefits for vulnerable candidates.

The Problem with Videos. Algorithmic screening can take advantage of data from multiple sources, including gameplay, psychological assessments, and video interviews. The latter provide three types of signals, namely visual, verbal, and paraverbal. Across multiple studies, systems trained on video signals are shown to yield the largest disparities and disadvantages for protected groups [33, 120, 276]. Visual signals have been removed from several products [175] because they inevitably encode sensitive information such as race and gender while lacking a solid foundation to justify their use in the hiring domain. Even if found to be accurate and fair in a specific evaluation, hiring algorithms based on face analysis are unlikely to predictably generalize and maintain accuracy or fairness under variable conditions, or at least they are less likely to do so than algorithms based on more established data modalities and better-understood correlations with job performance. Furthermore, it is worth noting regulation proposals against inference of emotions, states of mind, or intentions from face images in the workplace [84, 85]. Given this evidence, we invite particular caution against new proposals of hiring systems based on computer vision, advertised as capable of inferring the motivation [139] and personality traits [183] of candidates, even if accompanied by bias mitigation and fairness evaluation.

6 Data

Datasets used in algorithmic fairness research for the hiring domain are summarized in Table 4. The following sections present these resources divided into textual, visual, and tabular datasets. In line with recent literature, we find no graph dataset on algorithmic hiring [46, 71].

6.1 Text-Based Datasets

Textual datasets consist of job descriptions and job seekers' resumes or biographies, focusing on professional experience, training, and skills. Resumes and biographies contain strong proxies for sensitive attributes such as race and gender, including names and addresses.

Chinese bios is a textual dataset generated by Zhang et al. [283] to study gender bias in candidate rankings produced by BERT-based resume retrieval systems [70]. It consists of short resumes

Table 4. Datasets Used in Research on Fairness in Algorithmic Hiring

Dataset name	Used in	Type	Language	Geography	Target variable	Sensitive variable[a]	Hiring stage
Chinese Bios	[283]	Textual	Chinese	Synthetic	Mention of job-specific skills in bio	Gender (B, A)	Sourcing
Bias in Bios	[63, 119]	Textual	English	World	Candidate occupation	Gender (B, A)	Sourcing
Engineers and Scientists	[58]	Textual	English	Unknown	Candidate received an offer	Unknown	Screening
IT Resumes	[194, 196]	Textual	English	US	Candidate was called back	Gender (B, S)	Sourcing
CVs from Singapore	[69]	Textual	English	Singapore	CV match with job description	Ethnicity (A)	Sourcing
DPG Resumes	[222]	Textual	English, Dutch	The Netherlands	Candidate industry of interest	Gender (B, A)	Sourcing
ChaLearn First Impression	[120, 148, 276]	Visual	English	Unknown	Speaker hireability and personality annotated by AMT workers	Gender (B, A), ethnicity (A)	Screening
Student Interviews	[33]	Visual	English	Unknown	Speaker hireability annotated by research assistants	Gender (NB, S)	Screening
SHL Interviews	[231]	Visual	English	US, UK, India	Speaker communication skills rated by experts	Gender (B), age, race, country of residence	Screening
Oil Company Interviews	[139]	Visual	Dutch	The Netherlands	Candidate self-reported commitment	Gender, age	Screening
FairCVs	[119, 199]	Visual	English	Synthetic	Synthetic score	Gender (B, A), race (A)	Sourcing
Requirements and Candidates	[172]	Tabular	Synthetic	Synthetic	Synthetic score	Synthetic	Sourcing
Jobs and Candidates	[11]	Tabular	Unknown	The Netherlands	Candidate recruited or shortlisted	Gender (B, S)	Sourcing
Pymetrics Bias Group	[271]	Tabular	Unknown	Unknown	Similarity to current employees	Gender (S), race (S)	Screening
IBM HR Analytics	[105]	Tabular	Unknown	Unknown	Employee resignation	Gender (B), age, marital status	Evaluation
Resume Search Engines	[48]	Tabular	English	US	Unknown	Gender (B, S)	Sourcing
Walmart Employees	[38]	Tabular	English	US	Employee tenure and performance ratings	Synthetic	Screening
Web Developers Field Study	[13]	Tabular	English	US	Unknown	Gender (B,S), age (S), ethnicity (S)	Screening
Chinese Job Recommendations	[282]	Tabular	Chinese	China	Unknown	Gender (B,A), age (A)	Sourcing
Facebook Ads Audiences	[4]	Tabular	English	US	User clicks	Gender (B, S), race (B, A)	Sourcing

[a]Where reported, we indicate the provenance of sensitive attributes as annotated (A) or self-reported (S); we also indicate the gender encoding as binary (B) or non-binary (NB).

containing a binary gender indicator (he/she) and a two-sentence description of job skills for IT, finance, and administration.

Bias in bios [63] is composed of textual biographies written in English and extracted from the Common Crawl dataset. It was initially proposed to study fairness in occupation classification. Gender is automatically extracted based on the use of pronouns in the short biographies. Professions are the target variables; they are self-reported in the descriptions.

Engineers and scientists [58] results from a field study comparing human and algorithmic CV screening in an unspecified company. It is composed of applications, i.e., pairs of resumes and job postings for engineers (software and hardware) and technical scientists. The features were parsed from resumes, including candidate education (institutions, degrees, majors, awards), work experience (job titles and companies), skills, and relevant keywords. The target variable indicates whether the candidates were interviewed and extended an offer.

IT resumes [196] was used to study how men and women describe themselves on resumes and whether the difference impacts hiring outcomes. The dataset comprises approximately 900,000 resumes (without names, e-mails, and URLs) in the historical hiring records of eight IT firms based in the US, relevant to just over 6,000 job postings from the IT sector. The resumes were managed through an applicant tracking system, where the applicants self-reported their gender. The target variable encodes whether candidates received a callback after applying.

CVs from Singapore [69] was introduced to investigate ethnicity bias in automated resume filtering. It contains 135 resumes of candidates of Chinese, Malaysian, and Indian origin (the predominant ethnic groups in Singapore) who applied for vacancies in Singapore's accounting and finance sectors. The dataset curators annotated candidate ethnicity based on geographical information on their education and early employment. For example, candidates who completed their education in China were classified as ethnically Chinese. Nine job postings describing open positions in the financial sector are considered. Three annotators annotated each posting-resume pair with a binary variable indicating whether candidates appear qualified for a job based on their CV.

DPG resumes [222] contains over 10 million vacancies (including salary range, working hours, and job descriptions) and just under 1 million resumes augmented with job categories of interest and gender inferred from first names. Given the imbalance between vacancies and candidates, this dataset is suitable for studying recommender systems that propose jobs to candidates. The data are provided by DPG Recruitment in anonymized form, removing all names (including company names), dates, addresses, telephone numbers, e-mail addresses, and Web sites.

6.2 Visual Datasets

Several datasets in the hiring space are multimodal, with a strong focus on videos of candidates answering job interview questions in front of a camera [33, 139, 231], often called AVI. This data modality is relatively new in the hiring domain. Similarly to CVs, AVIs encode much information on sensitive attributes, including gender, race, age, and disability.

ChaLearn First Impression [79] contains 10,000 YouTube video clips of people facing a camera and speaking in English. Amazon Mechanical Turk workers were hired to annotate each clip with the personality traits of the speaker (openness, conscientiousness, extroversion, agreeableness, neuroticism) and a "variable indicating whether the subject should be invited to a job interview or not." The gender and ethnicity of the speakers were annotated by two dataset curators.

Student Interviews [33] consists of video interviews with 733 upper-level undergraduate students, who were asked to participate in a simulated interview answering six questions. Verbal (e.g., n-gram frequencies), paraverbal (e.g., loudness, jitter, shimmer), and visual (e.g., facial expressions, body motion) features were automatically extracted from videos using tools like OpenSmile [88], FACET [238], or Motion Tracker [270]. Data were annotated by three research assistants who assessed the candidates' "hireability" on a 5-point Likert scale. Gender is self-reported, including a non-binary option.

SHL Interviews [231] is a proprietary AVI dataset with more than 5,000 videos from 810 real job seekers from the US, UK, India, and Europe answering behavioral and domain knowledge questions. Videos were annotated by at least five assessors, who rated the presence of four social skills: engagement, positive emotion, calmness, and confidence, using a scale of 0 to 4. The sensitive attributes available with the dataset are country, age, gender, and race.

Oil Company Interviews [139] were curated to study the problem of predicting candidate motivation in job selection processes. It comprises AVIs with 154 students from Utrecht University carrying out a mock interview with a fictitious oil company. The participants self-reported their motivation ("To what degree are you motivated to work for the company") on a 10-point Likert scale, which represents the target variable. Software tools like OpenFace and EMFACS [75] were used to automatically create facial marker features and extract emotions from videos.

FairCVs [199] is a synthetic CV dataset combining short bios from *Bias in bios* [63], face images taken from the *DiveFace* database [180], and numerical features emulating desirable aspects such as availability or previous experience. Artificial target scores are generated for each CV, as a linear combination of numerical features; biased scores are derived from these with an ethnicity- and

gender-dependent additive penalty emulating biases in the data. The dataset has been employed to study the extent to which sensitive attribute proxies can contribute to discriminatory models when the target variables on which they are trained exhibit biases against certain protected groups.

6.3 Tabular Data

Tabular datasets encode structured data of a diverse nature, describing job seekers and employees at different stages of the algorithmic hiring pipeline.

Requirements and Candidates [172] is a synthetic dataset with numerical and Boolean values encoding both candidate skills and job requirements. Bias against certain candidates is deliberately introduced in the data through additive noise.

Jobs and Candidates [11] is a tabular dataset describing candidates and job postings with real-valued, categorical, and binary features. For candidates, their education, experience, preferences (minimum salary, preferred working hours, maximum travel distance), and self-reported gender are included. Regarding job postings, the dataset contains information about the industry, company size, and geographical location. Joint candidate-job features describing overlaps, such as the distance of candidates' residence from the place of work, are also present. The target variable indicates whether a candidate was recruited or short-listed for a job.

Pymetrics Bias Group [271] is a test set used for pre-deployment audits at Pymetrics, a company offering gameplay-based screening tools to clients. The gameplay of job seekers and current employees of the client company are compared to find candidates who resemble high-performing incumbent employees. The covariates consist of gamified psychological measurements; sensitive attributes, which can be self-reported on a voluntary basis, include gender and race.

IBM HR Analytics[1] is a synthetic dataset curated by IBM data scientists to study employee resignation, which is the target variable. Covariates include education, job satisfaction, income, years of service in the company, and commuting distance, along with sensitive attributes such as gender, age, and marital status.

Resume Search Engines [48] includes search results crawled from employment Web sites *Indeed*, *Monster*, and *CareerBuilder* for 35 job titles in 20 cities of the US, collecting data on 855,000 job candidates. Data were crawled in 2016 to study gender biases.

Walmart Employees was released as part of the Society for Industrial and Organizational Psychology machine learning competition[2] to study the problem of predicting employee retention and performance from pre-employment tests with questions on work history, personality, and behavioral scenarios. Target variables encode employee tenure and performance ratings. Each instance has a synthetic binary variable that mimics a protected attribute.

Web Developers Field Study [13] summarizes the results of an experiment on the impact of algorithmic hiring tools on gender diversity. The curators advertised a web developer position for US residents; upon applying, candidates provided information on their education, experience, and demographics, as well as free-form responses to selected questions. The responses were rated using a recruitment tool with a score of up to 100. Each applicant was then rated by a human assessor based on experience and education; assessors were divided into three groups based on whether they had access to algorithmic scores and candidate names.

Chinese Job Recommendations [282] consists of job recommendations from four Chinese boards to fictitious profiles that differ only by gender. Profiles are accessed programmatically every 2 weeks to record job recommendations, which are then compared between different genders.

[1] Available at https://github.com/IBM/employee-attrition-aif360
[2] https://eval.ai/web/challenges/challenge-page/527/overview

Several tabular datasets have been curated to measure discrimination in job ad delivery [4, 132, 158], using a common methodology. *Facebook Ads Audiences* [4] exemplifies this methodology by running a job ad campaign on Facebook and studying differential delivery along gendered and racial lines. They design advertising creatives (headline, text, and image) for different professions, using them in a campaign that optimizes clicks, which are then broken down into different demographics. Despite the fact that the Facebook Marketing API does not allow a breakdown by race, the curators perform this analysis with a careful design of target audiences leveraging phone numbers and racial information available from North Carolina voter records.

6.4 Discussion

Low Diversity. English is by far the dominant language; datasets with geographical information primarily represent US citizens (six datasets) and the Netherlands (three datasets). We offer two interpretations of these findings. On the one hand, they reflect the importance of both countries in the recruitment industry.[3] On the other hand, they are consistent with previous research reporting that most efforts to improve fairness in AI are influenced by the Global North [187, 221], particularly the US [135], and focus on English language resources [210].

Missing Stages. The vast majority of datasets describe the early stages of the hiring pipeline, i.e., sourcing (11 out of 20) and screening (8 out of 20). We found no datasets (and consequently no studies) for selection, and only one for evaluation (*IBM HR Analytics*). This is expected given the industry's tendency to use algorithms primarily for sourcing and screening [125, 166]. However, given the growing push to adopt this technology in later stages [259], there is a tangible risk that algorithms for selection and evaluation will be quietly deployed without a clear understanding of risks and limitations. Indeed, datasets that target employee tenure and performance, such as *Walmart Employees* and *IBM HR Analytics*, signal an interest from companies in understanding the factors that predict future productivity and loyalty.

Lacking Sensitive Attributes. Most importantly, attributes such as disability, religion, and sexual orientation are simply missing, despite the special legal status of these attributes and the evidence of workplace discrimination [7, 191, 193]. Gender is by far the most common sensitive attribute, overwhelmingly encoded as binary. Ethnicity and race are considered in 6 out of 20 datasets, making this the second most common attribute. They are frequently annotated, rather than self-reported, in textual and visual datasets, due to these data modalities carrying strong proxies for sensitive attributes.

Target Multiplicity. HR management and recruitment tasks allow multiple formulations. Many target variables can seem reasonable at face value; therefore, initial data curation and design choices have a prominent role. As candidates move through the hiring pipeline, their digital record goes through a data journey where they are marked as aware of a position, applying (or headhunted), proposed to a client, screened-in, interviewed, hired, retained, promoted, and so on. Companies interested in algorithmic hiring solutions pick one or more of these variables, balancing different priorities such as efficiency and quality of hire, with unpredictable effects on algorithmic fairness, a phenomenon called *multi-target multiplicity* [266]. Indeed, the target variables in Table 4 are very diverse. On the one hand, this reflects the length of hiring pipelines and the diversity of data journeys. On the other hand, it points to a lack of established best practices around target variables. Focusing on screening datasets, we find different constructs (e.g., communication skills vs. commitment) annotated by people with different competencies (AMT workers vs. experts) from disparate data sources (YouTube videos vs. mock interviews). Since the validity of these estimates

[3] https://www2.staffingindustry.com/eng/Editorial/Daily-News/World-World-s-largest-staffing-firms-post-224-billion-in-revenue-56012

and their connection with job performance has been called into question [19, 214], we call for caution in handling these variables and granting them the status of *ground truth*. This can be especially problematic when using conditional fairness measures whose very definition hinges on this so-called ground truth.

7 Measurement and Mitigation in Practice

In this section, we present practical considerations on bias mitigation and measurement that emerge from our review and from direct involvement in the industry. We focus on key technical factors guiding measurement and mitigation choices.

7.1 Data Modalities

First, different data modalities enable different bias mitigation strategies.

Textual data is a common data modality encountered in the sourcing stage, typically in the form of textual CV or job description data, or in the screening stage, e.g., transcripts of video interviews. Common bias mitigation methods that cater exclusively to textual data include rule-based scraping or dictionary-based methods. These methods, also offered by vendors such as *Textmetrics* and *Textio*, can be used for mitigating specific types of biases, e.g., age bias in job descriptions [98], or gender bias in biographies [63]. These mitigation strategies tend to be technically straightforward to implement, as they can be developed as standalone components and need not be integrated in complex algorithmic hiring pipelines. In addition, collecting the required data (e.g., constructing task-specific dictionaries) leans heavily on domain and context-specific knowledge, which is typically available to practitioners in the HR domain. A drawback of these dictionary-based methods is a different side of the same coin: due to reliance on domain, context, and task-specific data (dictionaries) they do not transfer well over different problems (e.g., scraping gendered words from resumes vs. substituting words associated with age discrimination in job descriptions require wholly different sets of terms), languages, or geographies (where cultural or regulatory differences may impose different requirements).

Additionally, when it comes to textual bias mitigation, the rapid uptake of LLMs has pulled into focus biases that arise from LLM-based natural language generation. For example, Salinas et al. [224] show how LLMs exhibit gender and nationality bias when generating job recommendations for job seekers of different genders and nationalities. They find that the types of jobs recommended follow common gender and nationality-based stereotypes. In a similar experiment, GPT-4 is found to more frequently use female pronouns in reference letters generated for female-dominated occupations (such as nannies), and male pronouns for male-dominated occupations (such as plumbers) [35]. A simple yet effective mitigation strategy here is prompt engineering: by appending the phrase "in an inclusive way" to the prompt, previously gendered pronouns are replaced with third person pronouns ("they/their"). While novel bias mitigation strategies for LLMs are actively studied [167], the previous examples show that being mindful of which information to include in a prompt (either explicit or implicit), and understanding how to effectively construct prompts are important first mitigation steps for leveraging LLMs in the context of algorithmic hiring.

Next, *visual data*, either through images or video, enable and require a different set of bias mitigation methods and strategies. As shown in Section 5, video-based systems tend to yield more disadvantageous effects for protected groups, also due to the strong encoding of sensitive information. In this light, *EASYRECRUE* present an adversarial method that removes sensitive information from latent representations of neural networks that were trained for predicting "hireability" given (features that represent) facial expressions of candidates in job interviews [120]. This type of adversarial bias mitigation can, however, be applied over a wider set of data modalities. In the end, due to video-based algorithms unreliability in the face of varying conditions, and incoming regulation

proposals that prohibit inference of emotions or intentions from facial data in workplace contexts (discussed in more detail in Section 5.4), it is advisable for practitioners to approach video-based hiring systems and tools with caution, irrespective of which mitigation strategy to choose.

Finally, *tabular data* is a common data modality in algorithmic hiring and machine learning more broadly. Mitigating bias in tabular data can be done through pre-processing approaches that directly intervene on the features, e.g., by removing features that are highly correlated with sensitive attributes via importance-based scraping, or increasing the representation of vulnerable populations in training sets with balanced sampling. These types of methods are widely available in open-source bias mitigation software packages such as the *Fairlearn* Python package [268] or the *AI Fairness 360* toolkit (available in Python and R) [21]. In addition, we see practitioners experiment with alternative pre-processing methods in algorithmic hiring, such as synthetic tabular data generation [197] for measuring [256] and mitigating bias [11].

7.2 Tasks

Next to different types of data modalities, different downstream tasks can enable and call for different measurement and bias mitigation strategies. The two most common tasks in algorithmic hiring are classification and ranking.

For classification tasks, getting started with bias mitigation is relatively straightforward, as there exist several open source or otherwise freely available software packages and libraries, as mentioned above, that provide different implementations of bias measurement methods and mitigation algorithms, specifically designed for classification tasks.

Ranking is a common task in the sourcing and screening stages of hiring, performed by search engines or recommender systems. However, the prevalence of classification-based bias metrics in algorithmic hiring (discussed in more detail in Section 4) is reflected in the aforementioned open source packages, which means that measuring and mitigating bias in ranking systems tends to follow the practice of re-purposing classification methods through, e.g., thresholding rankings (i.e., cutting off at k). We recommend using fair ranking measures for an evaluation of ranking systems that is more cognizant of user browsing behavior, although, of course, the latter needs to be suitably modeled. While several mitigation methods apply to both classification and ranking tasks (e.g., adversarial inference), practitioners should be aware of ranking-specific methods such as DetGreedy.

7.3 Scalability and Efficiency

Technical and infrastructural decisions further affect which bias mitigation methods to consider. For example, Geyik et al. [103] decouple their post-processing method from specific model choices and properties of input data, which means their method can naturally scale across the different (ranking) systems at *LinkedIn*. This also means that such a method can be developed and deployed as a standalone component or micro-service, which decreases development time, effort, and alignment, when compared to pre-processing or in-processing methods that may need to be designed and integrated in complex algorithmic hiring pipelines. In general, post-processing approaches offer an advantage in terms of system integration, but they should be applied with care due to anti-discrimination law (Section 8).

In terms of algorithmic efficiency, the re-ranking method by Geyik et al. [103] for bias mitigation can be considered computationally cheap as only a subset of a model's output needs to be processed (i.e., top-k items). At the same time, while low, the additional computational costs will be incurred for each ranker output. Bias mitigation through pre-processing, such as counterfactual data augmentation, and in-processing, such as feature weighting, can be comparatively more resource intensive as they involve (re-)training models. Adversarial inference methods, which have been applied to a

variety of data modalities and tasks, even require training multiple neural networks simultaneously, which means they can incur substantially higher costs compared to their non-mitigated counterparts. However, while these additional costs may be incurred at training time, the resulting models do not incur additional costs at inference time. In the end, algorithmic efficiency and computational costs will vary across bias mitigation methods in nature and magnitude, with simple rule-based scraping methods that involve dictionary look-ups on the cheap end of the spectrum, and bias mitigation methods that involve re-training LLMs at the resource-intensive end [167]. Aspects such as task, model complexity, architecture, training parameters, dataset size, and composition influence the scalability and efficiency of bias mitigation strategies.

7.4 Sensitive Attribute Data Availability and Usage

Different bias mitigation methods have different requirements around the availability of sensitive data, which can further steer mitigation method selection.

The availability of sensitive data is affected by several factors in practice. First, as we identify in Section 8, access to sensitive attributes may be at odds with privacy regulations such as the **General Data Protection Regulation (GDPR)**, which is an important real-world constraint. Next, availability of and access to sensitive attributes may require job seekers' explicit consent, which can be difficult to acquire at the scale needed for some bias mitigation methods. Finally, some sensitive attributes, such as age and gender, may have to be recorded in the hiring process for identification purposes and can hence be assumed to be available for all job seekers. Many other sensitive attributes (e.g., religious beliefs) will not be easily available and will have to be explicitly requested for bias measurement and mitigation purposes.

This lack of access to sensitive attributes means some bias mitigation methods may be less suited than others, as noted in Section 5.4; in particular, post-processing methods that require sensitive attributes for all candidates at inference time can be unrealistic. Here, pre-processing bias mitigation methods may prove more useful, as they (only) require access to sensitive attributes at training time and can operate on attributes even when available for only a subset of job seekers (e.g., those who provided consent). Furthermore, this allows these methods to be deployed and run in isolation from production systems in controlled batch scenarios, which can be another important practical benefit. Examples of these bias mitigation methods used by practitioners include adversarial inference [222], and re-balancing training data with synthetic data generation [11].

Finally, a third family of bias mitigation methods that can operate without requiring access to sensitive attributes at all are the rule-based approaches described above, which are commonly used for mitigating bias in textual data such as job descriptions or resumes. These methods rely solely on domain and task-specific gazetteers or dictionaries and hence can be used when access to sensitive attributes of individuals is not available or desirable.

7.5 Fairness Definitions and Intervention Targets

Once data modality, task, infrastructural and technical choices, and access to sensitive data are set, one important practical challenge is that of defining "fairness" and formulating the intervention target of a bias mitigation method—i.e., deciding "when to intervene" and "what to optimize for."

In the case of *LinkedIn*'s Talent Search [103], the DetGreedy algorithm was implemented to have the ranker's top-k output reflect the gender distribution of job seekers that meet the requirements of a recruiter-issued query, i.e., the desired distribution of the ranking is set to be equivalent to that of the underlying population of job seekers. Here, alignment with the company's goals and values, interpretability (or ability to explain), and collaboration across different stakeholders were mentioned as key factors in guiding the eventual target distribution. Defining fairness and formulating an intervention target is not a one-off task, as algorithmic hiring components will be

deployed in different, product-specific contexts and stages of the hiring pipeline. Candela et al. [39] present the (evolving) framework used at *LinkedIn* that guides definitions and operationalizations of AI fairness across their different (types of) products.

In general, the challenge in defining fairness and formulating intervention targets in algorithmic hiring is exacerbated by the multi-stakeholder nature of the hiring domain, where development teams may need to consult legal and compliance teams, HR professionals and recruiters, product managers, and executives, all of whom may need to be informed or provide input. Indeed, these choices should be guided by ethical, social, and legal dimensions. This challenge may explain in part the popularity of DI as a fairness metric, due to its ease of interpretation across a broad and diverse stakeholder group, or the popularity of complying with the **Equal Employment Opportunity Commission (EEOC)** 80% rule [77] as a fairness target (adopted, e.g., by *Pymetrics* [271]), as it appears, at least superficially, a legally grounded target.

7.6 Fairness vs. Utility Tradeoffs

In the sourcing or screening stages, depending on the intervention target, *outcome fairness* may come at the cost of utility; this can happen, for example, when optimizing for gender parity in heavily male- or female-dominated industries. However, perhaps surprisingly, different experiments by practitioners show that outcome fairness does not need to come at the cost of utility. First, in their online A/B-testing experiments, Geyik et al. [103] find their DetGreedy method improves their fairness metric, with no significant impact on utility. In addition, Arafan et al. [11] find that their pre-processing mitigation method of rebalancing training data may even improve utility over non-bias-mitigated methods, in an offline experiment using real data from an international HR company. Moreover, with respect to *impact fairness*, Peng et al. [200] show that "overcompensation" (i.e., artificially over-representing a gender in the output of a ranking system) as an intervention strategy for a hiring algorithm can in some cases mitigate human bias further down the hiring funnel.

Overall, this section described several practical considerations constraining bias mitigation interventions; desired utility levels are just one dimension, and often not the most restrictive.

8 Legal Landscape

In this section, we describe the main regulations and non-discrimination provisions concerning algorithmic hiring in the EU and the US. We list the main legal sources in both regions, emphasizing the former since it is less cited in the computer science literature on algorithmic hiring. We close this section with remarks on open challenges and practical concerns at the intersection of technology and policy.

8.1 European Law

Non-Discrimination Law. We focus on rules that apply in the whole European Union (27 Member States). In the absence of specific legal rules on non-discrimination in algorithmic hiring, general non-discrimination law applies. The right to non-discrimination is protected as a human right in Europe by the European Convention on Human Rights (1950) and the Charter of Fundamental Rights of the European Union (2020). The EU also adopted a number of legal acts called *directives*, prohibiting several types of discrimination in different contexts, which Member States implement (give effect to) by adopting national law [153]. The four most relevant non-discrimination directives are the following: the Racial Equality Directive (2000), the Employment Equality Directive (2000), the Gender Goods and Services Directive (2004), and the Recast Gender Equality Directive (2006) [54–56, 86]. Together, the directives offer protection against discrimination in hiring on the basis of six grounds, also called protected characteristics, or protected attributes: age; disability; gender;

religion or belief; racial or ethnic origin; and sexual orientation. EU law distinguishes two categories of prohibited discrimination: *direct* and *indirect*.

Direct discrimination means that the person is treated less favorably than another on the basis of a protected characteristic, such as ethnicity [54]. For example, if a company says it will not recruit people of a certain ethnicity, that is an example of direct discrimination [57]. Direct discrimination is always prohibited, with a few narrowly defined and specific examples. For instance, a women's clothing brand is allowed to hire only female models for its advertising pictures.

The second category of prohibited discrimination is *indirect discrimination*, occurring when a practice is neutral at first glance but ends up discriminating against people of a certain ethnicity (or another protected characteristic) [54]. For indirect discrimination, the law focuses on the effects of a practice; the intention of the alleged discriminator is not relevant. Hence, even if an organization can prove that it did not know that its algorithmic system discriminated unfairly, that will not help the organization.

There are three elements of indirect discrimination that can be summarized as follows [34, 54]. (1) The practice must be neutral. For example, rejecting job applications coming from a certain postal code would count as neutral. Rejecting applications from people with a certain ethnicity would not be neutral; it would be direct discrimination. (2) This neutral practice puts people with a certain ethnicity (or other sensitive attributes) at a "particular disadvantage compared with other persons." The word *disadvantage* must be interpreted in a wide way. (3) There is no objective justification for such practice. The apparently neutral practice is not prohibited if the "practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary."

As a hypothetical example, suppose that people with an immigrant background make more spelling errors in job application letters. A cleaning company never hires job applicants for cleaning jobs if the application contains spelling errors. This practice seems neutral at first glance, but results in the rejection of most applicants with an immigrant background. The cleaning company cannot justify this no-spelling-errors rule because people can be good cleaners, even if they make some spelling errors. Therefore, the cleaning company engages in illegal indirect discrimination. However, the situation would be different for a law firm. The main job of many lawyers is writing precise, and often official, documents. The law firm is allowed to reject applications with spelling errors, even if it results in most people with an immigrant background not being hired.

The organization is also responsible if it uses an algorithmic system provided by, for instance, a company. If the algorithmic system turns out to discriminate illegally, the organization using the system is responsible and the victim can sue it for damages, for instance. (Later, the organization could try to sue the AI developer, but the organization remains responsible toward the victim.) In sum, general non-discrimination law applies to new forms of algorithmic discrimination, also if that discrimination happens indirectly or accidentally.

Other Relevant Law. We briefly highlight some other relevant laws in the EU. The GDPR is, roughly summarized, a European-wide statute that aims to protect fairness and human rights when personal data are used. The GDPR is long and detailed. Among other things, GDPR bans the use of *special categories* of personal data (sometimes called sensitive data). These are data on, for example, someone's ethnicity, religion, trade union membership, health, or sexual orientation (article 9 GDPR). There are some exceptions to the ban. For instance, hospitals are allowed to use health data. Another exception is the individual's *explicit consent*. Generally speaking, consent is not freely given, and thus not valid, if an employer asks a job applicant or employee for their consent, because of the unequal power relation. This GDPR rule makes it difficult to use sensitive data to audit or train algorithmic systems [254].

There is a proposal for an AI Act in the EU, with many requirements for "high-risk" systems, including algorithmic systems for HR [84]. Developers of high-risk AI systems must, for instance, ensure that the training data are appropriate and do not lead to unlawful discrimination. At the time of writing, the EU did not officially adopt the AI Act yet, and did not publish the final text yet. The proposals in the AI Act also contain a new exception to the GDPR, to enable the use of sensitive data for debiasing algorithmic systems.

8.2 US Law

Federal Law. US non-discrimination law is similar to EU law in many respects, but also decidedly different in others. Like EU law, its sources are spread over different statutes, and case law plays a crucial role. For example, Title VII of the Civil Rights Act of 1964 constitutes a federal law prohibiting employment discrimination based on race, color, religion, sex, and national origin. Significantly, the EEOC issues non-binding, but practically important guidelines to interpret Title VII. Other important sources of federal law are the Age Discrimination in Employment Act of 1967, the Americans with Disabilities Act of 1990, and the Genetic Information Nondiscrimination Act of 2008.

For all these acts, US law distinguishes between two fundamental types of discrimination: disparate treatment and DI. They resemble, but do not perfectly mirror, the difference between direct and indirect discrimination in the EU. Importantly, disparate treatment requires not only an adverse action based on a protected attribute but also the proof of intent on the part of the discriminating individual—unlike the EU variety of direct discrimination [254]. To actually win in court, an injured person in the spelling mistake example would have to demonstrate that the cleaning agency introduced the spelling requirement with the purpose of treating unfavorably members of their protected group. In practice, this will often be difficult [251].

DI, in turn, closely resembles indirect discrimination. Intent is not required [249]. Rather, the DI doctrine prohibits actions that are seemingly neutral, but significantly disadvantage members of a protected group. In its guidelines, the EEOC suggested that such a disadvantage usually occurs if the chance of a member of the protected group being positively evaluated is 80% or less than that of a member of the privileged group (so-called 80% rule or 4/5 rule) [77]. Finally, DI can be justified if there is a legitimate reason for the practice [250], for example business necessity [18, 249]. Therefore, a law firm could arguably use a model evaluating orthography and grammar to rank candidates even if this disproportionately disadvantages members of one specific protected group.

Overall, unless intent can be shown, many cases of algorithmic discrimination will be argued under the DI prong. Therefore, many contributions to the literature on technical algorithmic fairness have provided tools to ensure that this rule is not violated at the statistical level [94, 278, 279]. However, courts will generally, both in the EU and in the US, look at factors beyond mere scores and numbers to determine whether a legally relevant disadvantage exists [112, 260].

Other Relevant Law. Apart from the acts mentioned, federal legislation specifically addressing discrimination in AI systems is unlikely to emerge anytime soon. The Algorithmic Accountability Act is stalled in a gridlocked Congress. Hence, several states and municipalities have taken the initiative and enacted AI hiring laws themselves. For example, the city of New York passed a law on automated employment decision tools [186]. From 5 July 2023, NYC Local Law 144 applies to employers using AI to substantially assist or replace discretionary decision-making in hiring. They are required to conduct and publish impartial bias evaluations by an independent auditor. Furthermore, the state of Illinois enacted the AI Video Interview Act [131]. In force since the year 2020, the law requires employers to put candidates on notice and obtain their consent, before

subjecting their video interview to AI analysis. Candidates may request deletion, and employers that rely solely on AI need to collect and report candidates' race and ethnicity.

Practitioners have to comply with these local rules if their model is applied in these jurisdictions. Finally, further constraints may arise from affirmative action law, particularly if the fairness intervention goes beyond what is necessary to remedy otherwise unjustified discrimination; this is a complex topic in both the US [22, 145, 252] and the EU legal framework [112, 126].

8.3 Operational Challenges

This outline of the EU and US legal landscapes provides some normative reference points for practitioners and offers an opportunity to discuss some practical challenges in assessing the legal compliance of algorithms.

In general terms, an algorithmic system does not cause indirect discrimination or DI if it pursues *legitimate aims* through *appropriate means*. In practice, several questions arise about both elements at the base of this legal principle. Is inferring a candidate's motivation, i.e., an internal combination of their emotions, states of mind, and intentions, a legitimate aim? Moreover, should estimates by algorithms trained on biased "ground truth" variables be considered appropriate means? In the absence of precise guidelines, these questions have to be assessed contextually to each algorithmic system, and algorithmic implementations should be compared, to the best extent possible, to (hypothetical) outcomes under non-algorithmic alternatives [112].

Direct discrimination, in turn, is generally much more difficult to legally justify [2]. A natural question arises about the compliance of bias mitigation approaches embedded in hiring algorithms. More specifically, are post-processing algorithms legitimate only if they enforce inter-group parity for equally qualified candidates, e.g., targeting TPR parity rather than DI? If so, what quantitative criteria should be applied to assess candidate qualifications? And to what extent is human intervention and a comprehensive assessment of any re-ranking required? Here, the answers depend on affirmative action (US) and positive action (EU) law, which are currently in flux after the US Supreme Court's decision against Harvard's and UNC's affirmative action programs. At a minimum, human oversight of post-processing operations, and an evaluation of its effects both on groups and on individuals particularly worthy of protection (single parents; chronically ill persons), seems advisable.

9 The Real World Is Messy

Our previous discussion has focused on how discrimination can be exacerbated by algorithms. Nevertheless, we have striven to point out the ways such discrimination can be mitigated and even how algorithms can be designed to actively discourage discrimination. It should now be clear that the design of technology plays a key role, either in creating discrimination or by reducing it. To overly value the role of technology design in discrimination or anti-discrimination would also be a mistake, however. This is because non-technological factors have been shown, on occasion, to strongly influence decisions and amplify technological bias, as we will explore in this section.

9.1 Algorithmic Uptake

During the COVID-19 pandemic, delivery companies such as Amazon, Deliveroo, and DoorDash rolled out algorithmic recruitment systems to avoid the danger of viral contagion in their HR teams, as well as experiment with a new technology that had the potential to save millions of dollars in HR bills [281]. One reason they were able to do this with ease was due to the regulatory environment, which was laxer than usual due to the ongoing emergency conditions. We have explored some of the ethical problems regarding discrimination in previous sections of this article, as well as some of the potential technological solutions to designing algorithms with anti-discrimination in mind.

Each of these solutions is a technological response, however, so we need to remain aware of the non-technological or reduced-technological options.

Although hiring technology can be more ethical than humans and reduce bias in decision-making [138], it often reinforces bias and results in distinctly non-ethical outcomes. This means we should at least consider the balance of ethical harms if we reintroduced a non-technological solution. The recent pandemic necessitated social distancing, but once this danger had passed, non-technological hiring procedures could be reintroduced. To understand why this has not happened, we must consider the institutional incentives of tech companies. These companies have invested heavily in their digital hiring technology as suppliers or customers, so are highly motivated to retain it, even once there is no longer a strong need for it.

9.2 Algorithmic Fairness

Recalling the example from Section 3.4, suppose that appropriate technical solutions have been deployed in hiring algorithms to counter the BCFs against intersectional minorities. Should she seek a new job, JS would have a sizeable probability of being sourced, screened-in, and recruited—equal to natives (male and non-male alike) with similar skills and experience in the hospitality industry. However, in the presence of an exogenous shock, such as the COVID-19 pandemic, workplaces are heavily affected, with large consequences on hiring and recruitment. The hospitality industry receives a major blow. Food services stop hiring waitstaff and dismiss most of the employees. Schools and childcare services become unavailable. All of a sudden, JS finds herself unemployed, urgently in need of a new job, which is now more difficult to obtain since demand for her skills has decreased and competition has surged. Sourcing algorithms place JS at the bottom positions of rankings, granting her low visibility. Childcare duties demand much time from her due to her gender [227]. She has fewer connections than native job seekers for recommendation and referral [284]. Overall, her probability of successfully reaching the end of a hiring pipeline has dropped substantially and more sharply than for privileged groups. In addition, her migration background makes JS less likely to gain support from social safety nets [45, 171], increasing the urgency of her need and the unfairness of the new *status quo*.

This section highlights two limitations of algorithmic fairness in hiring. On the one hand, by restricting their scope to a single system at a precise point in time, fairness measurements run the risk of missing the bigger picture, i.e., the broader socio-technical system in which hiring algorithms are embedded, which can change quickly and profoundly. A model deemed accurate and fair in the old context may perform poorly in the new one. On the other hand, these changes may be difficult to detect and quantify. Fairness evaluations by practitioners on pre-deployment test sets, such as *Pymetrics Bias Group* (Section 6.3), may quickly become obsolete. Strong shocks in the hiring domain are an issue for data representativeness more broadly. Fresh data become necessary. This is further complicated by changing incentive structures, by the complexity of handling sensitive data, and by the frequent delay between decisions and feedback in hiring.

10 Opportunities and Limitations

Algorithmic hiring benefits from, and contributes to, fairness and anti-discrimination work, as the previous sections have shown. In this section, we summarize the emerging opportunities and limitations, from which we derive a set of recommendations for researchers and practitioners, summarized in Table 5.

10.1 Bias and Validity

Opportunities. Algorithms for hiring can consider large pools of candidates, avoiding the preliminary exclusion of unusual profiles, as often done by human recruiters under time constraints.

Table 5. Opportunities, Limitations, and Recommendations for Algorithmic Hiring and Fairness Research

	Opportunities	Limitations	Recommendations
Bias and Validity	Consider large candidate pools, reduce human biases, and attract minority candidates	Risk of encoding individual biases along with inevitable societal biases; invalid target variables	Focus on vulnerable populations beyond acceptance rates; study individual fairness and exploratory policies; carefully scrutinize new technologies
Broader Context	Trigger positive feedback loops; consider tech-recruiter collaboration	Narrow focus on local outcomes can overlook fairness in entire hiring process; risk of repurposing for termination	Center on job seeker impacts; identify leaks in pipelines; design for recruiter-tech interaction; beware of performance and tenure prediction
Data	Support evaluations of diversity and inclusion	Reduced geographical, linguistic, and sensitive attribute coverage	Design data collection for diversity; develop IP-friendly audits
Law	Apply binding regulation to positively influence industry	Legal restrictions on fairness approaches: concerns about discrimination and data protection	Multidisciplinary research balancing fairness, privacy, and anti-discrimination; monitor EU AI Act

IP, intellectual property.

Under-representation and sampling biases can be mitigated as a result. Algorithmic hiring also has the potential to mitigate biases in imperfect human judgments. The simple fact of using algorithmic decision-making can reduce avoidance by vulnerable candidates [13]. Fair and trustworthy algorithms can lead to a positive form of automation bias and attract minority candidates.

Limitations. Data-driven algorithms tend to encode individual and societal biases. Some algorithms are explicitly trained and evaluated to "predict the competency scores candidates would have been given by trained human reviewers" [189], inheriting individual biases from recruiters. Previous experience is a preeminent feature for assessing candidates. In conjunction with current job segregation, this means that the most important features are inevitably skewed against historically disadvantaged groups. In addition to these biases, the epistemic validity of prediction targets such as candidates' *employability* and *motivation* is questionable. Job performance is famously difficult to define and measure, let alone predict [220, 263]. Algorithms cloaked in objectivity can promote bias while targeting and legitimizing ill-defined quantities.

Recommendations. Attention should be devoted not only to acceptance rates for vulnerable populations (DI) but also to their representation among applicants, as well as their progress downstream of the algorithm and post-hiring. Job descriptions and organizational communication can play an important role in attracting or repelling specific groups; automation attempts [207] should carefully include fairness evaluations. To mitigate biases against unusual candidates, individual fairness and exploratory policies should be studied. Individual fairness measures can surface problematic situations for individuals that may go unnoticed when studying group fairness. Exploratory policies based on partially stochastic mechanisms can provide new information in repeated decision-making scenarios. However, the social acceptability and procedural justice of such policies in the hiring domain remain to be studied. Finally, new technologies, including AVIs and personality prediction, deserve additional scrutiny, especially through the lens of validity theory [213, 214].

10.2 Broader Context

Opportunities. It is worth highlighting that improving fairness does not require completely removing bias. Algorithmic hiring can reduce certain disparities and trigger deflating feedback loops across

BCFs. These algorithms do not (and should not) operate autonomously. Effective and equitable hiring can result from a fruitful interaction between technology and recruiters, leveraging complementary strengths. For example, HR professionals are better suited to assess special cases and operate under changing conditions [166].

Limitations. Most fairness measures are focused on narrow algorithmic outcomes, neglecting the wider socio-technical context around these algorithms. Some of these measures are completely symmetric and consider advantages for vulnerable and privileged groups as equally problematic (which is, however, generally required by the law). Zooming out from single-algorithm evaluations, it is worth noting that outcome fairness at every stage does not guarantee fair outcomes for job seekers throughout the hiring pipeline. Furthermore, fairness for employers is currently missing; two-sided platforms would be well-positioned to study the performance of their algorithms across employers, devoting special attention to small businesses. Finally, it is worth noting that the discovery of patterns that predict job performance for hiring can open the way to models for termination decisions.

Recommendations. We call for contextualized and integrated evaluations of decision-making processes that go beyond the predictions of a single algorithm. This will help to address complex problems, such as identifying leaks in the hiring pipeline that are most critical for vulnerable groups and modeling impacts on job seekers, such as their efforts and benefits. To exemplify, rejected candidates may still benefit to some extent from a specific type of explanation. Moreover, it will be important to better understand the utility derived from these algorithms by different employers, considering recruitment workflows and developing new fairness measures. The prevalent *human vs. algorithm* evaluation framework is of limited utility; to overcome it, more research on recruiter-machine interaction is required [240], including candidate screening models [6] leading to more granular measures. Finally, we invite special caution in the development of predictive models for job performance and tenure, due to a risk of exploitation for termination decisions [285]; such an application of algorithms raises even stronger ethical and social concerns, which are only beginning to be discussed [223].

10.3 Data

Opportunities. Algorithmic fairness research is contributing additional analysis into hiring practices from a perspective of diversity and non-discrimination. More data entail more scrutiny and reflection, which can inform organizational frameworks, such as *Diversity, equity, and inclusion*, and scholarly fields, such as applied psychology and economics.

Limitations. Research on fairness in algorithmic hiring is based on data with reduced geographical and linguistic coverage. In addition, important sensitive attributes are missing from the data, making it difficult or impossible to evaluate algorithms for specific vulnerable groups. Data and research are constrained by a dual tension with the privacy of data subjects, on the one side, and the **Intellectual Property (IP)** of companies, on the other side.

Recommendations. Practitioners and researchers should seek more diverse data, with greater geographical and linguistic diversity, and better coverage of sensitive attributes that are relevant in hiring but are lacking, such as disability status and sexual orientation. Dedicated initiatives should be undertaken, including optional surveys for job applicants and broader data donation campaigns [25]. Innovative auditing protocols should be studied for employers and providers of algorithmic hiring solutions, including IP-friendly data disclosure procedures.

10.4 Law

Opportunities. In practice, binding regulation shapes algorithmic development more than ethical guidelines or self-regulation. Although clear guidance on algorithmic hiring is currently missing,

precise requirements set out in future regulation and case law, informed by research and practice on fair algorithmic hiring, have the potential to influence the industry positively and profoundly.

Limitations. Most of the fairness approaches developed so far are restricted to proxy reduction or removal, neglecting a wealth of solutions developed by the algorithmic fairness community. This is most likely due to concerns of infringing regulation on disparate treatment and direct discrimination. Furthermore, special categories of personal data are often lacking and difficult to process, particularly in the EU under the current data protection law. Therefore, it is difficult to assess hiring practices for certain vulnerable populations. Data protection law restricts these analyses and should consider exceptions for algorithmic fairness, as now suggested in the EU AI Act.

Recommendations. Algorithmic fairness can conflict with privacy and non-discrimination doctrine. The exact contours of legally compliant algorithmic fairness remain contested. The EU AI Act may offer a (limited) solution by allowing certain types of sensitive data processing to remove biases in high-risk scenarios. This guidance should be expanded to other areas and jurisdictions. We advocate for further multidisciplinary research on this topic, studying technical solutions and legal frameworks to reconcile these principles in light of their tradeoffs. Promising technological approaches include multiparty computation [143], sample-level estimators [89], and noise injection mechanisms [137].

11 Conclusions

The social, technological, and legal landscape around algorithmic hiring is rapidly evolving; algorithmic fairness has become a necessary component for both business-as-usual product development and frontier research. Practitioners and researchers in this field must understand BCFs, leveraging contextualized measures carried out on appropriate data to deploy suitable bias mitigation strategies. Multidisciplinary work at the intersection with legal scholarship is especially critical to implement and guide policy by defining technically achievable desiderata. Only a contextualized and balanced understanding of fair algorithmic hiring can guide research and practice to avoid the pitfalls of legitimizing questionable applications with misguided analyses and to reap truly shared benefits for society.

Acknowledgment

We are indebted to many researchers and practitioners for advice on this work, including Anisha Nadkarni, Anna Via, Carlos Castillo, Clara Rus, Didac Fortuny Almiñana, Feng Lu, Ilir Kola, Justine Devos, Marc Serra Vidal, and Volodymyr Medentsiy.

References

[1] Mohsen Abbasi, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Fairness in representation: Quantifying stereotyping as a representational harm. In *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM '19)*. Tanya Y. Berger-Wolf and Nitesh V. Chawla (Eds.), SIAM, 801–809. DOI: https://doi.org/10.1137/1.9781611975673.90

[2] Jeremias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth. 2023. Directly discriminatory algorithms. *The Modern Law Review* 86, 1 (2023), 144–175.

[3] Ifeoma Ajunwa. 2019. The paradox of automation as anti-bias intervention. *Cardozo Law Review* 41 (2019), 1671.

[4] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 199:1–199:30. DOI: https://doi.org/10.1145/3359301

[5] Kristen M. Altenburger, Rajlakshmi De, Kaylyn Frazier, Nikolai Avteniev, and Jim Hamilton. 2017. Are there gender differences in professional self-promotion? An empirical case study of LinkedIn profiles among recent MBA graduates. In *Proceedings of the 11th International Conference on Web and Social Media (ICWSM '17)*. AAAI Press, 460–463. Retrieved from https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15615

[6] Jose M. Alvarez and Salvatore Ruggieri. 2023. The initial screening order problem. arXiv:2307.15398. Retrieved from https://arxiv.org/abs/2307.15398

[7] Mason Ameri, Lisa Schur, Meera Adya, F. Scott Bentley, Patrick McKay, and Douglas Kruse. 2018. The disability employment puzzle: A field experiment on employer hiring behavior. *ILR Review* 71, 2 (2018), 329–364.

[8] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. The standards for educational and psychological testing.

[9] Lori Andrews and Hannah Bucher. 2022. Automating discrimination: AI hiring practices and gender inequality. *Cardozo Law Review* 44 (2022), 145.

[10] Julia Angwin, Noam Scheiber, and Ariana Tobin. 2017. *Dozens of companies are using facebook to exclude older workers from job ads*. Machine Bias. ProPublica, New York, NY. Retrieved from https://www.propublica.org/article/facebook-ads-age-discrimination-targeting

[11] Adam Mehdi Arafan, David Graus, Fernando P. Santos, and Emma Beauxis-Aussalet. 2022. End-to-end bias mitigation in candidate recommender systems with fairness gates. In *Proceedings of the 2nd Workshop on Recommender Systems for Human Resources (RecSys-in-HR '22)*. CEUR-WS, 1–8.

[12] Aditya Auradkar, Chavdar Botev, Shirshanka Das, Dave De Maagd, Alex Feinberg, Phanindra Ganti, Lei Gao, Bhaskar Ghosh, Kishore Gopalakrishna, Brendan Harris, Joel Koshy, Kevin Krawez, Jay Kreps, Shi Lu, Sunil Nagaraj, Neha Narkhede, Sasha Pachev, Igor Perisic, Lin Qiao, Tom Quiggle, Jun Rao, Bob Schulman, Abraham Sebastian, Oliver Seeliger, Adam Silberstein, Boris Shkolnik, Chinmay Soman, Roshan Sumbaly, Kapil Surlaker, Sajid Topiwala, Cuong Tran, Balaji Varadarajan, Jemiah Westerman, Zach White, David Zhang, and Jason Zhang. 2012. Data infrastructure at LinkedIn. In *Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE '12)*. Anastasios Kementsietsidis and Marcos Antonio Vaz Salles (Eds.), IEEE Computer Society, 1370–1381. DOI: https://doi.org/10.1109/ICDE.2012.147

[13] Mallory Avery, Andreas Leibbrandt, and Joseph Vecci. 2023. Does artificial intelligence help or hurt gender diversity? Evidence from two field experiments on recruitment in tech. Retrieved from http://monash-econ-wps.s3.amazonaws.com/RePEc/mos/moswps/2023-09.pdf

[14] Chen Avin, Barbara Keller, Zvi Lotker, Claire Mathieu, David Peleg, and Yvonne-Anne Pignolet. 2015. Homophily and the glass ceiling effect in social networks. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science (ITCS '15)*. Tim Roughgarden (Ed.), ACM, 41–50. DOI: https://doi.org/10.1145/2688073.2688097

[15] Ghazala Azmat and Barbara Petrongolo. 2014. Gender and the labor market: What have we learned from field and lab experiments? *Labour Economics* 30 (2014), 32–40.

[16] Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM* 61, 6 (2018), 54–61. DOI: https://doi.org/10.1145/3209581

[17] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. Retrieved from http://www.fairmlbook.org

[18] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671–732.

[19] Murray R. Barrick and Michael K. Mount. 1991. The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology* 44, 1 (1991), 1–26.

[20] Arthur G. Bedeian, Gerald R. Ferris, and K. Michele Kacmar. 1992. Age, tenure, and job satisfaction: A tale of two perspectives. *Journal of Vocational Behavior* 40, 1 (1992), 33–48.

[21] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. Retrieved from https://arxiv.org/abs/1810.01943

[22] Jason R. Bent. 2019. Is algorithmic affirmative action legal. *The Georgetown Law Journal* 108 (2019), 803.

[23] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94, 4 (2004), 991–1013.

[24] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2212–2220.

[25] Matthew Bietz, Kevin Patrick, and Cinnamon Bloss. 2019. Data donation as a model for citizen science health research. *Citizen Science: Theory and Practice* 4, 1 (2019).

[26] Lee E. Biggerstaff, Joanna T. Campbell, and Bradley A. Goldie. 2023. Hitting the "Grass Ceiling": Golfing CEOs, exclusionary schema, and career outcomes for female executives. *Journal of Management* 50 (2023), 1502–1535.

[27] Su Lin Blodgett, Lisa Green, and Brendan T. O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural*

Language Processing (EMNLP '16). Jian Su, Xavier Carreras, and Kevin Duh (Eds.), The Association for Computational Linguistics, 1119–1130. DOI: https://doi.org/10.18653/v1/d16-1120

[28] Lotte Bloksgaard. 2011. Masculinities, femininities and work–the horizontal gender segregation in the Danish Labour market. *Nordic Journal of Working Life Studies* 1, 2 (2011), 5–21.

[29] Donna Bobbitt-Zeher. 2011. Gender discrimination at work: Connecting gender stereotypes, institutional policies, and gender composition of workplace. *Gender & Society* 25, 6 (2011), 764–786.

[30] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*. Technical Report. Upturn.

[31] Jasmijn C. Bol. 2011. The determinants and performance effects of managers' performance evaluation biases. *The Accounting Review* 86, 5 (2011), 1549–1575.

[32] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*. Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.), 4349–4357. Retrieved from https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[33] Brandon M. Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K. D'Mello. 2021. Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the International Conference on Multimodal Interaction (ICMI '21)*. Zakia Hammal, Carlos Busso, Catherine Pelachaud, Sharon L. Oviatt, Albert Ali Salah, and Guoying Zhao (Eds.), ACM, 268–277. DOI: https://doi.org/10.1145/3462244.3479897

[34] Frederik Zuiderveen Borgesius. 2020. Price discrimination, algorithmic decision-making, and European non-discrimination law. *European Business Law Review* 31, 3 (2020), 401–422.

[35] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712.

[36] Pawan Budhwar, Ashish Malik, M. T. Thedushika De Silva, and Praveena Thevisuthan. 2022. Artificial intelligence–challenges and opportunities for international HRM: A review and research agenda. *The International Journal of Human Resource Management* 33, 6 (2022), 1065–1097.

[37] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT '18)*. Sorelle A. Friedler and Christo Wilson (Eds.), Vol. 81, PMLR, 77–91. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a.html

[38] Ian Burke, Robin Burke, and Goran Kuljanin. 2021. Fair candidate ranking with spatial partitioning: Lessons from the SIOP ML competition. In *Proceedings of the 1st Workshop on Recommender Systems for Human Resources (RecSys in HR '21) Co-located with the 15th ACM Conference on Recommender Systems (RecSys '21)*, Vol. 2967.

[39] Joaquin Quiñonero Candela, Yuwen Wu, Brian Hsu, Sakshi Jain, Jennifer Ramos, Jon Adams, Robert Hallman, and Kinjal Basu. 2023. Disentangling and operationalizing AI fairness at LinkedIn. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, 1213–1228. DOI: https://doi.org/10.1145/3593013.3594075

[40] David Card, Ana Rute Cardoso, and Patrick Kline. 2016. Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women. *The Quarterly Journal of Economics* 131, 2 (2016), 633–686.

[41] Ben Carterette. 2011. System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '11)*. Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft (Eds.), ACM, 903–912. DOI: https://doi.org/10.1145/2009916.2010037

[42] Census Bureau. 2023. Current population survey. Retrieved from https://stats.bls.gov/news.release/empsit.toc.htm

[43] Maura Cerioli, Maurizio Leotta, and Filippo Ricca. 2020. What 5 million job advertisements tell us about testing: A preliminary empirical investigation. In *Proceedings of the 35th ACM/SIGAPP Symposium on Applied Computing*. Chih-Cheng Hung, Tomás Cerný, Dongwan Shin, and Alessio Bechini (Eds.), ACM, 1586–1594. DOI: https://doi.org/10.1145/3341105.3373961

[44] Simon Chandler. 2018. The AI ChatBot will hire you now. Retrieved from https://www.wired.com/story/the-ai-chatbot-will-hire-you-now/

[45] Lei Che, Haifeng Du, and Kam Wing Chan. 2020. Unequal pain: A sketch of the impact of the Covid-19 pandemic on migrants' employment in China. *Eurasian Geography and Economics* 61, 4–5 (2020), 448–463.

[46] April Chen, Ryan A. Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. 2023. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data* 18 (2023), 1–23.

[47] Jie Chen, Chunxia Zhang, and Zhendong Niu. 2018. A two-step resume information extraction algorithm. *Mathematical Problems in Engineering* 2018 (2018), 5761287.

[48] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.), ACM, 651. DOI: https://doi.org/10.1145/3173574.3174225

[49] Ruey-Cheng Chen, Qingyao Ai, Gaya Jayasinghe, and W. Bruce Croft. 2019. Correcting for recency bias in job recommendation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (Eds.), ACM, 2185–2188. DOI: https://doi.org/10.1145/3357384.3358131

[50] Sapna Cheryan, Allison Master, and Andrew N. Meltzoff. 2015. Cultural stereotypes as gatekeepers: Increasing girls' interest in computer science and engineering by diversifying stereotypes. *Frontiers in Psychology* 6 (2015), 49.

[51] Raj Chetty, David J. Deming, and John N. Friedman. 2023. Diversifying society's leaders? The causal effects of admission to highly selective private colleges. Working Paper 31492. National Bureau of Economic Research. DOI: https://doi.org/10.3386/w31492

[52] T. Anne Cleary. 1968. Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement* 5, 2 (1968), 115–124.

[53] David A. Cotter, Joan M. Hermsen, Seth Ovadia, and Reeve Vanneman. 2001. The glass ceiling effect. *Social Forces* 80, 2 (2001), 655–681.

[54] Council of the European Union. 2000. Council Directive 2000/43/EC implementing the principle of equal treatment between persons irrespective of racial or ethnic origin. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0043

[55] Council of the European Union. 2000. Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078

[56] Council of the European Union. 2004. Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32004L0113

[57] Court of Justice of the European Union. 2008. Centrum voor gelijkheid van kansen en voor racismebestrijding v Firma Feryn NV. Retrieved from https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A62007CJ0054

[58] Bo Cowgill. 2018. *Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resume Screening*. Columbia Business School, Columbia University, Vol. 29.

[59] Terry-Ann Craigie. 2020. Ban the box, convictions, and public employment. *Economic Inquiry* 58, 1 (2020), 425–445.

[60] Nick Craswell, Onno Zoeter, Michael J. Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM '08)*. Marc Najork, Andrei Z. Broder, and Soumen Chakrabarti (Eds.), ACM, 87–94. DOI: https://doi.org/10.1145/1341531.1341545

[61] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112. DOI: https://doi.org/10.1515/popets-2015-0007

[62] Skylar Davidson. 2016. Gender inequality: Nonbinary transgender people in the workplace. *Cogent Social Sciences* 2, 1 (2016), 1236511.

[63] Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. danah boyd and Jamie H. Morgenstern (Eds.), ACM, 120–128. DOI: https://doi.org/10.1145/3287560.3287572

[64] Jenny Yang Deirdre Mulligan. 2023. Hearing from the American people: How are automated tools being used to surveil, monitor, and manage workers? Retrieved from https://www.whitehouse.gov/ostp/news-updates/2023/05/01/hearing-from-the-american-people-how-are-automated-tools-being-used-to-surveil-monitor-and-manage-workers/

[65] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '19)*. Computer Vision Foundation/IEEE, 4690–4699. DOI: https://doi.org/10.1109/CVPR.2019.00482

[66] Megan Denver. 2020. Criminal records, positive credentials and recidivism: Incorporating evidence of rehabilitation into criminal background check employment decisions. *Crime & Delinquency* 66, 2 (2020), 194–218.

[67] Anne-Sophie Deprez-Sims and Scott B. Morris. 2010. Accents in the workplace: Their effects during a job interview. *International Journal of Psychology* 45, 6 (2010), 417–426.

[68] Ellora Derenoncourt, Chi Hyun Kim, Moritz Kuhn, and Moritz Schularick. 2022. *Wealth of Two Nations: The US Racial Wealth Gap, 1860-2020*. Technical Report. National Bureau of Economic Research.

[69] Ketki V. Deshpande, Shimei Pan, and James R. Foulds. 2020. Mitigating demographic bias in AI-based resume filtering. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*. Tsvi Kuflik, Ilaria Torre, Robin Burke, and Cristina Gena (Eds.), ACM, 268–275. DOI: https://doi.org/10.1145/3386392.3399569

[70] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

[71] Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. 2023. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering* 35, 10 (2023), 10583–10602. DOI: https://doi.org/10.1109/TKDE.2023.3265598

[72] Alice H. Eagly, Christa Nater, David I. Miller, Michèle Kaufmann, and Sabine Sczesny. 2019. Gender stereotypes have changed: A cross-temporal meta-analysis of US public opinion polls from 1946 to 2018. *American Psychologist* 75, 3 (2019), 301–315.

[73] Harrison Edwards and Amos J. Storkey. 2016. Censoring representations with an adversary. In *Proceedings of the 4th International Conference on Learning Representations (ICLR '16)*. Yoshua Bengio and Yann LeCun (Eds.). arXiv:1511.05897. Retrieved from http://arxiv.org/abs/1511.05897

[74] EEOC - US Equal Employment Opportunity Commission. 2023. Select issues: Assessing adverse impact in software, algorithms, and artificial intelligence used in employment selection procedures under title VII of the Civil Rights Act of 1964. Retrieved from https://www.eeoc.gov/select-issues-assessing-adverse-impact-software-algorithms-and-artificial-intelligence-used

[75] Paul Ekman and Wallace V. Friesen. 2003. Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Vol. 10. ISHK.

[76] Naomi Ellemers. 2018. Gender Stereotypes. *Annual Review of Psychology* 69, 1 (2018), 275–298.

[77] Equal Employment Opportunity Commission. 2015. Uniform guidelines on employment selection procedures.

[78] Stefan Eriksson and Jonas Lagerström. 2012. The labor market consequences of gender differences in job search. *Journal of Labor Research* 33 (2012), 303–327.

[79] Hugo Jair Escalante, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yağmur Güçlütürk, Umut Güçlü, Xavier Baró, Isabelle Guyon, Julio C. S. Jacques Junior, Meysam Madadi, Evelyne Viegas, Furkan Gürpinar, Achmadnoer Sukma Wicaksana, Cynthia C. S. Liem, Marcel A. J. van Gerven, and Rob van Lier. 2020. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing* 13, 2 (2020), 894–911.

[80] Ben Eubanks. 2022. *Artificial Intelligence for HR: Use AI to Support and Develop a Successful Workforce*. Kogan Page Publishers.

[81] European Commission. 2020. The gender pay gap situation in the EU. Retrieved from https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/gender-equality/equal-pay/gender-pay-gap-situation-eu_en

[82] European Institute for Gender Equality. 2020. Gender Equality Index 2020. Retrieved from https://eige.europa.eu/publications/gender-equality-index-2020-key-findings-eu

[83] European Institute for Gender Equality. 2023. Gender Equality Index. Retrieved from https://eige.europa.eu/gender-equality-index/2022/domain/work

[84] European Parliament. 2021. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[85] European Parliament. 2023. Artificial Intelligence Act: Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Retrieved from https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf

[86] European Parliament and Council of the European Union. 2004. Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation (recast). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32006L0054

[87] Christine L. Exley and Judd B. Kessler. 2022. The gender gap in self-promotion. *The Quarterly Journal of Economics* 137, 3 (2022), 1345–1381.

[88] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, 1459–1462.

[89] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. 2023. Measuring fairness under unawareness of sensitive attributes: A quantification-based approach. *Journal of Artificial Intelligence Research* 76 (2023), 1117–1180. DOI: https://doi.org/10.1613/jair.1.14033

[90] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2074–2152. DOI: https://doi.org/10.1007/s10618-022-00854-z

[91] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Tackling documentation debt: A survey on algorithmic fairness datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*. ACM, 2:1–2:13. DOI: https://doi.org/10.1145/3551624.3555286

[92] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing and Management* 57, 6 (2020), 102377. DOI: https://doi.org/10.1016/j.ipm.2020.102377

[93] Lídia Farré. 2016. Parental leave policies and gender equality: A survey of the literature. *Studies of Applied Economics* 34, 1 (2016), 45–60.

[94] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams (Eds.), ACM, 259–268. DOI: https://doi.org/10.1145/2783258.2783311

[95] Elena Fernández-del Río, Linda Koopmans, Pedro J. Ramos-Villagrasa, and Juan R. Barrada. 2019. Assessing job performance using brief self-report scales: The case of the individual work performance questionnaire. *Revista de Psicología del Trabajo y de las Organizaciones* 35, 3 (2019), 195–205.

[96] Benjamin Fish, Ashkan Bashardoust, danah boyd, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Gaps in information access in social networks? In *The World Wide Web Conference (WWW '19)*. Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.), ACM, 480–490. DOI: https://doi.org/10.1145/3308558.3313680

[97] Riccardo Fogliato, Alice Xiang, Zachary C. Lipton, Daniel Nagin, and Alexandra Chouldechova. 2021. On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.), ACM, 100–111. DOI: https://doi.org/10.1145/3461702.3462538

[98] A. S. Fokkens, C. J. Beukeboom, and E. Maks. 2018. *Leeftijdsdiscriminatie in vacatureteksten: Een geautomatiseerde inhoudsanalyse naar verboden leeftijd-gerelateerd taalgebruik in vacatureteksten: Rapport in opdracht van het College voor de Rechten van de Mens.*

[99] World Economic Forum. 2021. Human-centred artificial intelligence for human resources: A toolkit for human resources professionals. Retrieved from https://www3.weforum.org/docs/WEF_Human_Centred_Artificial_Intelligence_for_Human_Resources_2021.pdf

[100] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. arXiv:1609.07236. Retrieved from https://arxiv.org/abs/1609.07236

[101] Joseph Fuller, Manjari Raman, Eva Sage-Gavin, and Kristen Hines. 2021. *Hidden Workers: Untapped Talent.* Technical Report. Harvard Business School.

[102] Danielle Gaucher, Justin Friesen, and Aaron C. Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology* 101, 1 (2011), 109.

[103] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 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 (KDD '19)*. Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.), ACM, 2221–2231. DOI: https://doi.org/10.1145/3292500.3330691

[104] Sahin Cem Geyik, Qi Guo, Bo Hu, Cagri Ozcaglar, Ketan Thakkar, Xianren Wu, and Krishnaram Kenthapadi. 2018. Talent search and recommendation systems at LinkedIn: Practical challenges and lessons learned. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.), ACM, 1353–1354. DOI: https://doi.org/10.1145/3209978.3210205

[105] Azin Ghazimatin, Matthäus Kleindessner, Chris Russell, Ziawasch Abedjan, and Jacek Golebiowski. 2022. Measuring fairness of rankings under noisy sensitive information. In *Proceedings of the 2022 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 2263–2279. DOI: https://doi.org/10.1145/3531146.3534641

[106] Kate Goddard, Abdul V. Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127. DOI: https://doi.org/10.1136/amiajnl-2011-000089

[107] Diego Gómez-Zará, Leslie A. DeChurch, and Noshir S. Contractor. 2020. A taxonomy of team-assembly systems: Understanding how people use technologies to form teams. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 181:1–181:36. DOI: https://doi.org/10.1145/3415252

[108] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '19) (Long and Short Papers). Jill Burstein, Christy Doran, and Thamar Solorio (Eds.), Vol. 1, Association for Computational Linguistics, 609–614. DOI: https://doi.org/10.18653/v1/n19-1061

[109] Carlos Gradín. 2021. Occupational gender segregation in post-apartheid South Africa. *Feminist Economics* 27, 3 (2021), 102–133.

[110] Kelsey Gray, Angela Neville, Amy H. Kaji, Mary Wolfe, Kristine Calhoun, Farin Amersi, Timothy Donahue, Tracy Arnell, Benjamin Jarman, Kenji Inaba, Marc Melcher, Jon B. Morris, Brian Smith, Mark Reeves, Jeffrey Gauvin, Edgardo S. Salcedo, Richard Sidwell, Kenric Murayama, Richard Damewood, V. Prasad Poola, Daniel Dent, and Christian de Virgilio. 2019. Career goals, salary expectations, and salary negotiation among male and female general surgery residents. *JAMA Surgery* 154, 11 (2019), 1023–1029.

[111] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, Vol. 1, 11.

[112] Philipp Hacker. 2018. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law. *Common Market Law Review* 55, 4 (2018), 1143–1185.

[113] Aniko Hannak, Claudia Wagner, David García, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from TaskRabbit and Fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Charlotte P. Lee, Steven E. Poltrock, Louise Barkhuus, Marcos Borges, and Wendy A. Kellogg (Eds.), ACM, 1914–1933. DOI: https://doi.org/10.1145/2998181.2998327

[114] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 29 (2016), 3315–3323.

[115] Joyce C. He and Sonia K. Kang. 2021. Covering in cover letters: Gender and self-presentation in job applications. *Academy of Management Journal* 64, 4 (2021), 1097–1126.

[116] Joyce C. He, Sonia K Kang, Kaylie Tse, and Soo Min Toh. 2019. Stereotypes at work: Occupational stereotypes predict race and gender segregation in the workforce. *Journal of Vocational Behavior* 115 (2019), 103318.

[117] Deepak Hegde, Alexander Ljungqvist, and Manav Raj. 2022. Race, glass ceilings, and lower pay for equal work. *Swedish House of Finance Research Paper* 21, 09.

[118] Madeline E. Heilman. 2012. Gender stereotypes and workplace bias. *Research in Organizational Behavior* 32 (2012), 113–135.

[119] Léo Hemamou and William Coleman. 2022. Delivering fairness in human resources AI: Mutual information to the rescue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL/IJCNLP '22)*. Yulan He, Heng Ji, Yang Liu, Sujian Li, Chia-Hui Chang, Soujanya Poria, Chenghua Lin, Wray L. Buntine, Maria Liakata, Hanqi Yan, Zonghan Yan, Sebastian Ruder, Xiaojun Wan, Miguel Arana-Catania, Zhongyu Wei, Hen-Hsen Huang, Jheng-Long Wu, Min-Yuh Day, Pengfei Liu, and Ruifeng Xu (Eds.), Long Papers, Vol. 1. Association for Computational Linguistics, 867–882. Retrieved from https://aclanthology.org/2022.aacl-main.64

[120] Léo Hemamou, Arthur Guillon, Jean-Claude Martin, and Chloé Clavel. 2021. Don't judge me by my face: An indirect adversarial approach to remove sensitive information from multimodal neural representation in asynchronous job video interviews. In *Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction (ACII '21)*. IEEE, 1–8. DOI: https://doi.org/10.1109/ACII52823.2021.9597443

[121] Morela Hernandez, Derek R. Avery, Sabrina D. Volpone, and Cheryl R. Kaiser. 2019. Bargaining while black: The role of race in salary negotiations. *Journal of Applied Psychology* 104, 4 (2019), 581.

[122] Iñigo Hernandez-Arenaz and Nagore Iriberri. 2019. A review of gender differences in negotiation. In *Oxford Research Encyclopedia of Economics and Finance*.

[123] Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the moral justification of statistical parity. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT '21 '21)*. Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.), ACM, 747–757. DOI: https://doi.org/10.1145/3442188.3445936

[124] HireVue. 2022. Explainability statement. Retrieved from https://hirevue-api.dev-directory.com/wp-content/uploads/2022/04/HV_AI_Short-Form_Explainability_1pager.pdf

[125] Bilal Hmoud and Varallyai Laszlo. 2019. Will artificial intelligence take over human resources recruitment and selection. *Network Intelligence Studies* 7, 13 (2019), 21–30.

[126] Holly Hoch, Corinna Hertweck, Michele Loi, and Aurelia Tamò. 2021. Discrimination for the sake of fairness: Fairness by design and its legal framework. Available at SSRN 3773766.

[127] Md Sajjad Hosain, Ping Liu, and Mohitul Ameen Ahmed Mustafi. 2021. Social networking information and pre-employment background check: Mediating effects of perceived benefit and organizational branding. *International Journal of Manpower* 42 (2021).

[128] Shenggang Hu, Jabir Alshehabi Al-Ani, Karen D. Hughes, Nicole Denier, Alla Konnikov, Lei Ding, Jinhan Xie, Yang Hu, Monideepa Tarafdar, Bei Jiang, Linglong Kong, and Hongsheng Dai. 2022. Balancing gender bias in job advertisements with text-level bias mitigation. *Frontiers Big Data* 5 (2022), 805713. DOI: https://doi.org/10.3389/fdata.2022.805713

[129] David J. Hughes and Mark Batey. 2017. Using personality questionnaires for selection. In *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection and Employee Retention*, 151–181.

[130] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. danah boyd and Jamie H. Morgenstern (Eds.), ACM, 49–58. DOI: https://doi.org/10.1145/3287560.3287600

[131] Illinois General Assembly. 2020. Artificial Intelligence Video Interview Act, 820 ILCS 42. Retrieved from https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015 & ChapterID=68

[132] Basileal Imana, Aleksandra Korolova, and John S. Heidemann. 2021. Auditing for discrimination in algorithms delivering job ads. In *The World Wide Web Conference 2021 (WWW '21)*. Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.), ACM / IW3C2, 3767–3778. DOI: https://doi.org/10.1145/3442381.3450077

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

[134] Jilana Jaxon, Ryan F. Lei, Reut Shachnai, Eleanor K. Chestnut, and Andrei Cimpian. 2019. The acquisition of gender stereotypes about intellectual ability: Intersections with race. *Journal of Social Issues* 75, 4 (2019), 1192–1215.

[135] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (01 Sep. 2019), 389–399. DOI: https://doi.org/10.1038/s42256-019-0088-2

[136] Jobvite. 2021. *2021 Recruiter Nation Report*. Technical Report. Retrieved from https://www.jobvite.com/lp/2021-recruiter-nation-report/

[137] Marc Juarez and Aleksandra Korolova. 2023. "You can't fix what you can't measure": Privately measuring demographic performance disparities in federated learning. In *Workshop on Algorithmic Fairness through the Lens of Causality and Privacy*. PMLR, 67–85.

[138] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Hachette UK.

[139] Mitchel Kappen and Marnix Naber. 2021. Objective and bias-free measures of candidate motivation during job applications. *Scientific Reports* 11, 1 (2021), 21254.

[140] Navroop Kaur and Sandeep K. Sood. 2017. A game theoretic approach for an IoT-based automated employee performance evaluation. *IEEE Systems Journal* 11, 3 (2017), 1385–1394. DOI: https://doi.org/10.1109/JSYST.2015.2469102

[141] Heysem Kaya, Furkan Gürpinar, and Albert Ali Salah. 2017. Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs. In *Proceedings of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017*. IEEE Computer Society, 1651–1659. DOI: https://doi.org/10.1109/CVPRW.2017.210

[142] Nicolas Kayser-Bril. 2023. *LinkedIn Automatically Rates "Out-of-Country" Candidates as "Not Fit" in Job Applications*. Technical Report. AlgorithmWatch.

[143] Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. 2018. Blind justice: Fairness with encrypted sensitive attributes. In *Proceedings of the 35th International Conference on Machine Learning (ICML '18)*. Jennifer G. Dy and Andreas Krause (Eds.), Vol. 80, PMLR, 2635–2644. Retrieved from http://proceedings.mlr.press/v80/kilbertus18a.html

[144] Pauline T. Kim. 2016. Data-driven discrimination at work. *William & Mary Law Review* 58 (2016), 857.

[145] Pauline T. Kim. 2022. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *California Law Review* 110 (2022), 1539.

[146] Nicholas J. Klein and Michael J. Smart. 2017. Car today, gone tomorrow: The ephemeral car in low-income, immigrant and minority families. *Transportation* 44 (2017), 495–510.

[147] Patrick Kline, Evan K. Rose, and Christopher R. Walters. 2022. Systemic discrimination among large US employers. *The Quarterly Journal of Economics* 137, 4 (2022), 1963–2036.

[148] Alina Köchling, Shirin Riazy, Marius Claus Wehner, and Katharina Simbeck. 2021. Highly accurate, but still discriminatory: A fairness evaluation of algorithmic video analysis in the recruitment context. *Business & Information Systems Engineering* 63 (2021), 39–54.

[149] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13, 3 (2020), 795–848.

[150] Alla Konnikov, Nicole Denier, Yang Hu, Karen D. Hughes, Jabir Alshehabi Al-Ani, Lei Ding, Irina Rets, and Monideepa Tarafdar. 2022. BIAS Word inventory for work and employment diversity,(in) equality and inclusivity (Version 1.0). SocArXiv (2022). Retrieved from https://ideas.repec.org/p/osf/socarx/t9v3a.html

[151] Margaret Bull Kovera. 2019. Racial disparities in the criminal justice system: Prevalence, causes, and a search for solutions. *Journal of Social Issues* 75, 4 (2019), 1139–1164.
[152] Kurt Kraiger and J. Kevin Ford. 1985. A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology* 70, 1 (1985), 56.
[153] Jasper Krommendijk and Frederik Zuiderveen Borgesius. 2023. EU law analysis 'How to read EU legislation?' Retrieved from http://eulawanalysis.blogspot.com/p/how-to-read-eu-legislation.html
[154] Deepak Kumar, Tessa Grosz, Navid Rekabsaz, Elisabeth Greif, and Markus Schedl. 2023. Fairness of recommender systems in the recruitment domain: An analysis from technical and legal perspectives. *Frontiers in Big Data* 6 (2023).
[155] Astrid Kunze and Amalia R. Miller. 2017. Women helping women? Evidence from private sector data on workplace hierarchies. *Review of Economics and Statistics* 99, 5 (2017), 769–775.
[156] Cordula Kupfer, Rita Prassl, Jürgen Fleiß, Christine Malin, Stefan Thalmann, and Bettina Kubicek. 2023. Check the box! How to deal with automation bias in AI-based personnel selection. *Frontiers in Psychology* 14 (2023), 1118723.
[157] Sarah E. Lageson, Elizabeth Webster, and Juan R. Sandoval. 2021. Digitizing and disclosing personal data: The proliferation of state criminal records on the internet. *Law & Social Inquiry* 46, 3 (2021), 635–665.
[158] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science* 65, 7 (2019), 2966–2981. DOI: https://doi.org/10.1287/mnsc.2018.3093
[159] Maude Lavanchy, Patrick Reichert, Jayanth Narayanan, and Krishna Savani. 2023. Applicants' fairness perceptions of algorithm-driven hiring procedures. *Journal of Business Ethics* 188 (2023), 1–26.
[160] M. Asher Lawson, Ashley E. Martin, Imrul Huda, and Sandra C. Matz. 2022. Hiring women into senior leadership positions is associated with a reduction in gender stereotypes in organizational language. *Proceedings of the National Academy of Sciences* 119, 9 (2022), e2026443119.
[161] Thomas Le Barbanchon, Roland Rathelot, and Alexandra Roulet. 2021. Gender differences in job search: Trading off commute against wage. *The Quarterly Journal of Economics* 136, 1 (2021), 381–426.
[162] Yeonjung Lee and Fengyan Tang. 2015. More caregiving, less working: Caregiving roles and gender difference. *Journal of Applied Gerontology* 34, 4 (2015), 465–483.
[163] Andreas Leibbrandt and John A. List. 2015. Do women avoid salary negotiations? Evidence from a large-scale natural field experiment. *Management Science* 61, 9 (2015), 2016–2024.
[164] Chee Wee Leong, Katrina Roohr, Vikram Ramanarayanan, Michelle P. Martin-Raugh, Harrison Kell, Rutuja Ubale, Yao Qian, Zydrune Mladineo, and Laura McCulla. 2019. Are humans biased in assessment of video interviews? In *Adjunct of the 2019 International Conference on Multimodal Interaction (ICMI '19)*. ACM, 9:1–9:5. DOI: https://doi.org/10.1145/3351529.3360653
[165] Eve A. Levin. 2018. Gender-normed physical-ability tests under Title VII. *Columbia Law Review* 118, 2 (2018), 567–604.
[166] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic hiring in practice: Recruiter and HR Professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 166–176.
[167] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. arXiv:2308.10149.
[168] Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil K. Jain, and Jiliang Tang. 2023. Trustworthy AI: A computational perspective. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2023), 4:1–4:59. DOI: https://doi.org/10.1145/3546872
[169] Yuxi Long, Jiamin Liu, Ming Fang, Tao Wang, and Wei Jiang. 2018. Prediction of employee promotion based on personal basic features and post features. In *Proceedings of the International Conference on Data Processing and Applications (ICDPA '18)*. ACM, 5–10. DOI: https://doi.org/10.1145/3224207.3224210
[170] Ashish Malik, Pawan Budhwar, Charmi Patel, and N. R. Srikanth. 2022. May the bots be with you! Delivering HR cost-effectiveness and individualised employee experiences in an MNE. *The International Journal of Human Resource Management* 33, 6 (2022), 1148–1178.
[171] Don Mar, Paul Ong, Tom Larson, and James Peoples. 2022. Racial and ethnic disparities in who receives unemployment benefits during COVID-19. *SN Business & Economics* 2, 8 (2022), 102.
[172] Karla Markert, Afrae Ahouzi, and Pascal Debus. 2022. Fairness in regression – Analysing a job candidates ranking system. In *INFORMATIK 2022.* Daniel Demmler, Daniel Krupka, and Hannes Federrath (Eds.), Gesellschaft für Informatik, 1275–1285. DOI: https://doi.org/10.18420/inf2022_109
[173] Yoosof Mashayekhi, Nan Li, Bo Kang, Jefrey Lijffijt, and Tijl De Bie. 2022. A challenge-based survey of e-recruitment recommendation systems. arXiv:2209.05112. Retrieved from https://arxiv.org/abs/2209.05112
[174] David A. Matsa and Amalia R. Miller. 2011. Chipping away at the glass ceiling: Gender spillovers in corporate leadership. *American Economic Review* 101, 3 (2011), 635–639.

[175] Roy Maurer. 2021. HireVue discontinues facial analysis screening. Retrieved from https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/hirevue-discontinues-facial-analysis-screening.aspx

[176] Qingxin Meng, Keli Xiao, Dazhong Shen, Hengshu Zhu, and Hui Xiong. 2022. Fine-grained job salary benchmarking with a nonparametric Dirichlet process-based latent factor model. *INFORMS Journal on Computing* 34, 5 (2022), 2443–2463. DOI: https://doi.org/10.1287/ijoc.2022.1182

[177] Alex Miller. 2018. Want less-biased decisions? Use algorithms. Retrieved from https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

[178] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[179] Tara Sophia Mohr. 2014. Why women don't apply for jobs unless they're 100% qualified. Retrieved from https://hbr.org/2014/08/why-women-dont-apply-for-jobs-unless-theyre-100-qualified

[180] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. 2020. SensitiveNets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 6 (2020), 2158–2164.

[181] Gordon B. Moskowitz, Jeff Stone, and Amanda Childs. 2012. Implicit stereotyping and medical decisions: Unconscious stereotype activation in practitioners' thoughts about African Americans. *American Journal of Public Health* 102, 5 (2012), 996–1001.

[182] Corinne A. Moss-Racusin and Laurie A. Rudman. 2010. Disruptions in women's self-promotion: The backlash avoidance model. *Psychology of Women Quarterly* 34, 2 (2010), 186–202.

[183] Dena F. Mujtaba and Nihar R. Mahapatra. 2021. Multi-task deep neural networks for multimodal personality trait prediction. In *Proceedings of the 2021 International Conference on Computational Science and Computational Intelligence (CSCI '21)*. IEEE, 85–91.

[184] Ann L. Mullen. 2009. Elite destinations: Pathways to attending an Ivy League university. *British Journal of Sociology of Education* 30, 1 (2009), 15–27.

[185] Preetam Nandy, Cyrus DiCiccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. 2022. Achieving fairness via post-processing in web-scale recommender systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, 715–725. DOI: https://doi.org/10.1145/3531146.3533136

[186] New York City Council. 2021. Automated employment decision tools, 144. Retrieved from https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced&Search

[187] Chinasa T. Okolo, Nicola Dell, and Aditya Vashistha. 2022. Making AI explainable in the global south: A systematic review. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS '22)*, 439–452.

[188] Oracle. 2023. Welcome to Oracle AI Apps for Talent Management. Retrieved from https://docs.oracle.com/en/cloud/saas/talent-management/22d/faimh/welcome-to-ai-apps-for-talent-management.html#u30010414

[189] ORCAA. 2020. *Description of Algorithmic Audit: Pre-built Assessments*. Technical Report. Retrieved from https://techinquiry.org/HireVue-ORCAA.pdf

[190] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 35, 27730–27744.

[191] Emir Ozeren. 2014. Sexual orientation discrimination in the workplace: A systematic review of literature. *Procedia-Social and Behavioral Sciences* 109 (2014), 1203–1215.

[192] KerryAnn O'Meara, Dawn Culpepper, and Lindsey L. Templeton. 2020. Nudging toward diversity: Applying behavioral design to faculty hiring. *Review of Educational Research* 90, 3 (2020), 311–348.

[193] Aasim I. Padela, Huda Adam, Maha Ahmad, Zahra Hosseinian, and Farr Curlin. 2016. Religious identity and workplace discrimination: A national survey of American Muslim physicians. *AJOB Empirical Bioethics* 7, 3 (2016), 149–159.

[194] Prasanna Parasurama and João Sedoc. 2021. Degendering resumes for fair algorithmic resume screening. arXiv:2112.08910. Retrieved from https://arxiv.org/abs/2112.08910

[195] Prasanna Parasurama and João Sedoc. 2022. Gendered information in resumes and its role in algorithmic and human hiring bias. In *Academy of Management Proceedings*, Vol. 2022. Academy of Management Briarcliff, Manor, NY, 17133.

[196] Prasanna Parasurama, João Sedoc, and Anindya Ghose. 2022. Gendered information in resumes and hiring bias: A predictive modeling approach. Available at SSRN 4074976.

[197] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA '16)*, 399–410. DOI: https://doi.org/10.1109/DSAA.2016.49

[198] Mirjana Pejic-Bach, Tine Bertoncel, Maja Mesko, and Zivko Krstic. 2020. Text mining of industry 4.0 job advertisements. *International Journal of Information Management* 50 (2020), 416–431. DOI: https://doi.org/10.1016/j.ijinfomgt.2019.07.014

[199] Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fiérrez. 2020. Bias in multimodal AI: Testbed for fair automatic recruitment. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020*. Computer Vision Foundation/IEEE, 129–137. DOI: https://doi.org/10.1109/CVPRW50498.2020.00022

[200] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What you see is what you get? The impact of representation criteria on human bias in hiring. In *Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP '19)*. Edith Law and Jennifer Wortman Vaughan (Eds.). AAAI Press, 125–134. Retrieved from https://ojs.aaai.org/index.php/HCOMP/article/view/5281

[201] Anders Persson. 2016. Implicit bias in predictive data profiling within recruitments. In *Privacy and Identity Management. Facing up to Next Steps - 11th IFIP WG* 9.2, 9.5, 9.6/11.7, 11.4, 11.6/SIG 9.2.2 International Summer School. Anja Lehmann, Diane Whitehouse, Simone Fischer-Hübner, Lothar Fritsch, and Charles D. Raab (Eds.), IFIP Advances in Information and Communication Technology, Vol. 498, 212–230. DOI: https://doi.org/10.1007/978-3-319-55783-0_15

[202] Dana Pessach, Gonen Singer, Dan Avrahami, Hila Chalutz Ben-Gal, Erez Shmueli, and Irad Ben-Gal. 2020. Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support System* 134 (2020), 113290. DOI: https://doi.org/10.1016/j.dss.2020.113290

[203] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, and Sharad Goel. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour* 4, 7 (2020), 736–745.

[204] Rohit Punnoose and Pankaj Ajit. 2016. Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence* 5, 9 (2016). DOI: https://doi.org/10.14569/IJARAI.2016.050904

[205] Sharon L, Segrest Purkiss, Pamela L. Perrewé, Treena L. Gillespie, Bronston T. Mayes, and Gerald R. Ferris. 2006. Implicit sources of bias in employment interview judgments and decisions. *Organizational Behavior and Human Decision Processes* 101, 2 (2006), 152–167.

[206] PwC. 2017. Artificial Intelligence in HR: A no-brainer. Retrieved from https://www.pwc.nl/nl/assets/documents/artificial-intelligence-in-hr-a-no-brainer.pdf

[207] Chuan Qin, Kaichun Yao, Hengshu Zhu, Tong Xu, Dazhong Shen, Enhong Chen, and Hui Xiong. 2023. Towards automatic job description generation with capability-aware neural networks. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2023), 5341–5355. DOI: https://doi.org/10.1109/TKDE.2022.3145396

[208] Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen. 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences* 114, 41 (2017), 10870–10875.

[209] Manish Raghavan, Solon Barocas, Jon M. Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.), ACM, 469–481. DOI: https://doi.org/10.1145/3351095.3372828

[210] Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics (EACL '23)*. Association for Computational Linguistics, Dubrovnik, Croatia, 2106–2119. Retrieved from https://aclanthology.org/2023.findings-eacl.157

[211] Christine Reyna, Mark Brandt, and G. Tendayi Viki. 2009. Blame it on hip-hop: Anti-rap attitudes as a proxy for prejudice. *Group Processes & Intergroup Relations* 12, 3 (2009), 361–380.

[212] Cecil R. Reynolds and Lisa A. Suzuki. 2012. Bias in psychological assessment: An empirical review and recommendations. In *Handbook of Psychology (2nd. ed.)*, Vol. 10.

[213] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, and Julia Stoyanovich. 2022. Resume format, LinkedIn URLs and other unexpected influences on AI personality prediction in hiring: Results of an audit. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Vincent Conitzer, John Tasioulas, Matthias Scheutz, Ryan Calo, Martina Mara, and Annette Zimmermann (Eds.), ACM, 572–587. DOI: https://doi.org/10.1145/3514094.3534189

[214] Alene K. Rhea, Kelsey Markey, Lauren D'Arinzo, Hilke Schellmann, Mona Sloane, Paul Squires, Falaah Arif Khan, and Julia Stoyanovich. 2022. An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Mining and Knowledge Discovery* 36, 6 (2022), 2153–2193.

[215] Peter A Riach and Judith Rich. 2002. Field experiments of discrimination in the market place. *The Economic Journal* 112, 483 (2002), F480–F518.

[216] Judith Rich. 2014. What do field experiments of discrimination in markets tell us? A meta analysis of studies conducted since 2000. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2517887

[217] Jonas Rieskamp, Lennart Hofeditz, Milad Mirbabaie, and Stefan Stieglitz. 2023. Approaches to improve fairness when deploying AI-based algorithms in hiring - using a systematic literature review to guide future research. In *Proceedings*

of the 56th Hawaii International Conference on System Sciences (HICSS '23). Tung X. Bui (Ed.), ScholarSpace, 216–225. Retrieved from https://hdl.handle.net/10125/102654

[218] Lauren A. Rivera. 2011. Ivies, extracurriculars, and exclusion: Elite employers' use of educational credentials. *Research in Social Stratification and Mobility* 29, 1 (2011), 71–90.

[219] Lauren A. Rivera. 2012. Hiring as cultural matching: The case of elite professional service firms. *American Sociological Review* 77, 6 (2012), 999–1022.

[220] David Robotham and Richard Jubb. 1996. Competences: Measuring the unmeasurable. *Management Development Review* 9, 5 (1996), 25–29.

[221] Cathy Roche, Dave Lewis, and P. J. Wall. 2021. Artificial intelligence ethics: An inclusive global discourse? arXiv:2108.09959. Retrieved from https://arxiv.org/abs/2108.09959

[222] Clara Rus, Jeffrey Luppes, Harrie Oosterhuis, and Gido H. Schoenmacker. 2022. Closing the gender wage gap: Adversarial fairness in job recommendation. In *Proceedings of the 2nd Workshop on Recommender Systems for Human Resources (RecSys-in-HR 2022) co-located with the 16th ACM Conference on Recommender Systems (RecSys 2022)*, Seattle, Washington, 1–10.

[223] Mary-Ann Russon. 2020. Uber sued by drivers over 'automated robo-firing'. *BBC News* 26 (2020).

[224] Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Examining demographic biases in job recommendations by ChatGPT and LLaMA. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. ACM. DOI: https://doi.org/10.1145/3617694.3623257

[225] Kimberly T. Schneider, Suzanne Swan, and Louise F. Fitzgerald. 1997. Job-related and psychological effects of sexual harassment in the workplace: Empirical evidence from two organizations. *Journal of Applied Psychology* 82, 3 (1997), 401.

[226] Frederike Scholz. 2020. Taken for granted: Ableist norms embedded in the design of online recruitment practices. In *The Palgrave Handbook of Disability at Work*, 451–469.

[227] Almudena Sevilla and Sarah Smith. 2020. Baby steps: The gender division of childcare during the COVID-19 pandemic. *Oxford Review of Economic Policy* 36, Supplement_1 (2020), S169–S186.

[228] Michael A. Shields and Stephen Wheatley Price. 2002. Racial harassment, job satisfaction and intentions to quit: Evidence from the British nursing profession. *Economica* 69, 274 (2002), 295–326.

[229] Jim Sidanius and Marie Crane. 1989. Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology* 19, 2 (1989), 174–197.

[230] Jan Simson, Alessandro Fabris, and Christoph Kern. 2024. Lazy data practices harm fairness research. In *Proceeding of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, 642–659. DOI: https://doi.org/10.1145/3630106.3658931

[231] Abhishek Singhania, Abhishek Unnam, and Varun Aggarwal. 2020. Grading video interviews with fairness considerations. arXiv:2007.05461. Retrieved from https://arxiv.org/abs/2007.05461

[232] Clea Skopeliti. 2023. 'I feel constantly watched': The employees working under surveillance. Retrieved from https://www.theguardian.com/money/2023/may/30/i-feel-constantly-watched-employees-working-under-surveillance-monitorig-software-productivity

[233] David G. Smith, Judith E. Rosenstein, Margaret C. Nikolov, and Darby A. Chaney. 2019. The power of language: Gender, status, and agency in performance evaluations. *Sex Roles* 80 (2019), 159–171.

[234] Daryl G. Smith, Caroline S. Turner, Nana Osei-Kofi, and Sandra Richards. 2004. Interrupting the usual: Successful strategies for hiring diverse faculty. *The Journal of Higher Education* 75, 2 (2004), 133–160.

[235] Lawrence B. Solum. 2004. Procedural justice. *Southern California Law Review* 78 (2004), 181.

[236] Keith E. Sonderling, Bradford J. Kelley, and Lance Casimir. 2022. The promise and the peril: Artificial intelligence and employment discrimination. *University of Miami Law Review* 77 (2022), 1.

[237] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28, 1 (1972), 11–21.

[238] Sabrina Stöckli, Michael Schulte-Mecklenbeck, Stefan Borer, and Andrea C. Samson. 2018. Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior Research Methods* 50 (2018), 1446–1460.

[239] Ana-Andreea Stoica, Christopher J. Riederer, and Augustin Chaintreau. 2018. Algorithmic glass ceiling in social networks: The effects of social recommendations on network diversity. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web (WWW '18)*. Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (Eds.). ACM, 923–932. DOI: https://doi.org/10.1145/3178876.3186140

[240] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2021. Does fair ranking improve minority outcomes? Understanding the interplay of human and algorithmic biases in online hiring. In *AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event (AIES '21)*. Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, 989–999. DOI: https://doi.org/10.1145/3461702.3462602

[241] Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the 1st ACL Workshop on Ethics in Natural Language Processing (EthNLP@EACL '17)*. Dirk Hovy, Shannon L. Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna M. Wallach (Eds.), Association for Computational Linguistics, 53–59. DOI: https://doi.org/10.18653/v1/w17-1606

[242] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson R. Murphy-Hill, Chris Parnin, and Jon Stallings. 2017. Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science* 3 (2017), e111. DOI: https://doi.org/10.7717/peerj-cs.111

[243] Rebbeca Tesfai and Kevin J. A. Thomas. 2020. Dimensions of inequality: Black immigrants' occupational segregation in the United States. *Sociology of Race and Ethnicity* 6, 1 (2020), 1–21.

[244] Kerri A. Thompson. 2020. Countenancing employment discrimination: Facial recognition in background checks. *Texas A & M Law Review* 8 (2020), 63.

[245] Nicholas Tilmes. 2022. Disability, fairness, and algorithmic bias in AI recruitment. *Ethics and Information Technology* 24, 2 (2022), 21. DOI: https://doi.org/10.1007/s10676-022-09633-2

[246] Shari Trewin, Sara H. Basson, Michael J. Muller, Stacy M. Branham, Jutta Treviranus, Daniel M. Gruen, Daniel Hebert, Natalia Lyckowski, and Erich Manser. 2019. Considerations for AI fairness for people with disabilities. *AI Matters* 5, 3 (2019), 40–63. DOI: https://doi.org/10.1145/3362077.3362086

[247] James M. Tyler and Jennifer Dane McCullough. 2009. Violating prescriptive stereotypes on job resumes: A self-presentational perspective. *Management Communication Quarterly* 23, 2 (2009), 272–287.

[248] UNDP - United Nations Development Programme. 2023. Breaking down gender biases: Shifting social norms towards gender equality. Retrieved from https://hdr.undp.org/system/files/documents/hdp-document/gsni202302pdf.pdf

[249] U.S. Supreme Court. 1971. Griggs v. Duke Power Co., 401 U.S. 424. Retrieved from https://supreme.justia.com/cases/federal/us/401/424/

[250] U.S. Supreme Court. 1973. McDonnell Douglas Corp. v. Green, 411 U.S. 792. Retrieved from https://supreme.justia.com/cases/federal/us/411/792/

[251] U.S. Supreme Court. 1989. Price Waterhouse v. Hopkins, 490 U.S. 228. Retrieved from https://supreme.justia.com/cases/federal/us/490/228/

[252] U.S. Supreme Court. 2009. Ricci v. DeStefano, 557 U.S. 557. Retrieved from https://supreme.justia.com/cases/federal/us/557/557/

[253] Chris Vallance. 2023. TUC: Government failing to protect workers from AI. Retrieved from https://www.bbc.com/news/technology-65301630

[254] Marvin Van Bekkum and Frederik Zuiderveen Borgesius. 2023. Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception? *Computer Law & Security Review* 48 (2023), 105770.

[255] Elmira van den Broek, Anastasia V. Sergeeva, and Marleen Huysman. 2019. Hiring algorithms: An ethnography of fairness in practice. In *Proceedings of the 40th International Conference on Information Systems (ICIS '19)*. Helmut Krcmar, Jane Fedorowicz, Wai Fong Boh, Jan Marco Leimeister, and Sunil Wattal (Eds.), Association for Information Systems. Retrieved from https://aisel.aisnet.org/icis2019/future_of_work/future_work/6

[256] Sarah-Jane van Els, David Graus, and Emma Beauxis-Aussalet. 2022. Improving fairness assessments with synthetic data: A practical use case with a recommender system for human resources. In *Proceedings of the 1st International Workshop on Computational Jobs Marketplace* (CompJobs '22), 5 pages.

[257] Sriram Vasudevan and Krishnaram Kenthapadi. 2020. LiFT: A scalable framework for measuring fairness in ML applications. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*. Mathieu d'Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux (Eds.), ACM, 2773–2780. DOI: https://doi.org/10.1145/3340531.3412705

[258] Giridhari Venkatadri and Alan Mislove. 2020. On the potential for discrimination via composition. In *ACM Internet Measurement Conference (IMC '20)*. ACM, 333–344. DOI: https://doi.org/10.1145/3419394.3423641

[259] Pranshu Verma. 2023. AI is starting to pick who gets laid off. Retrieved from https://www.washingtonpost.com/technology/2023/02/20/layoff-algorithms/

[260] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *Computer Law & Security Review* 41 (2021), 105567.

[261] Sean Waite. 2021. Should I stay or should I go? Employment discrimination and workplace harassment against transgender and other minority employees in Canada's federal public service. *Journal of Homosexuality* 68, 11 (2021), 1833–1859.

[262] Joseph Walker. 2012. Meet the new boss: Big Data. Retrieved from https://www.wsj.com/articles/SB10000872396390443890304578006252019616768

[263] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2022. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *Available at SSRN*.

[264] Yu Wang and Tyler Derr. 2022. Degree-related bias in link prediction. In *IEEE International Conference on Data Mining Workshops (ICDM '22)*. K. Selçuk Candan, Thang N. Dinh, My T. Thai, and Takashi Washio (Eds.), IEEE, 757–758. DOI: https://doi.org/10.1109/ICDMW58026.2022.00103

[265] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems* 41, 3 (2023), 52:1–52:43. DOI: https://doi.org/10.1145/3547333

[266] Jamelle Watson-Daniels, Solon Barocas, Jake M. Hofman, and Alexandra Chouldechova. 2023. Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, 297–311. DOI: https://doi.org/10.1145/3593013.3593998

[267] Amy L. Wax. 2011. Disparate impact realism. *William & Mary Law Review* 53 (2011), 621.

[268] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and improving fairness of AI systems, 8 pages. Retrieved from http://jmlr.org/papers/v24/23-0389.html

[269] Doris Weichselbaumer and Rudolf Winter-Ebmer. 2005. A meta-analysis of the international gender wage gap. *Journal of Economic Surveys* 19, 3 (2005), 479–511.

[270] Jacqueline Kory Westlund, Sidney K. D'Mello, and Andrew M. Olney. 2015. Motion tracker: Camera-based monitoring of bodily movements using motion silhouettes. *PloS One* 10, 6 (2015), e0130293.

[271] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proccedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 666–677. DOI: https://doi.org/10.1145/3442188.3445928

[272] Claes Wohlin. 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 1–10.

[273] Carol Woodhams, Ben Lupton, and Marc Cowling. 2015. The presence of ethnic minority and disabled men in feminised work: Intersectionality, vertical segregation and the glass escalator. *Sex Roles* 72 (2015), 277–293.

[274] Alison T. Wynn and Shelley J. Correll. 2018. Puncturing the pipeline: Do technology companies alienate women in recruiting sessions? *Social Studies of Science* 48, 1 (2018), 149–164.

[275] Renzhe Xu, Peng Cui, Kun Kuang, Bo Li, Linjun Zhou, Zheyan Shen, and Wei Cui. 2020. Algorithmic decision making with conditional fairness. In *Proccedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)*. ACM, 2125–2135. DOI: https://doi.org/10.1145/3394486.3403263

[276] Shen Yan, Di Huang, and Mohammad Soleymani. 2020. Mitigating biases in multimodal personality assessment. In *International Conference on Multimodal Interaction (ICMI '20)*. ACM, 361–369. DOI: https://doi.org/10.1145/3382507.3418889

[277] Maya Yaneva. 2018. Employee satisfaction vs. employee engagement vs. employee NPS. *European Journal of Economics and Business Studies* 4, 1 (2018), 221–227.

[278] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS '17)*, , Vol. 54. PMLR, 962–970. Retrieved from http://proceedings.mlr.press/v54/zafar17a.html

[279] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2020. Matching code and law: Achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery* 34, 1 (2020), 163–200. DOI: https://doi.org/10.1007/s10618-019-00658-8

[280] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2023. Fairness in ranking, part I: Score-based ranking. *ACM Computing Surveys* 55, 6 (2023), 118:1–118:36. DOI: https://doi.org/10.1145/3533379

[281] Lixuan Zhang and Christopher Yencha. 2022. Examining perceptions towards hiring algorithms. *Technology in Society* 68 (2022), 101848.

[282] Shuo Zhang and Peter Kuhn. 2022. Understanding algorithmic bias in job recommender systems: An audit study approach. (2022).

[283] Sijing Zhang, Ping Li, and Ziyan Cai. 2022. Are male candidates better than females? Debiasing BERT resume retrieval system. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC '22)*. IEEE, 616–621. DOI: https://doi.org/10.1109/SMC53654.2022.9945184

[284] Yiguang Zhang and Augustin Chaintreau. 2021. Unequal opportunities in multi-hop referral programs. arXiv:2112.00269. Retrieved from https://arxiv.org/abs/2112.00269

[285] Dave Zielinski. 2023. Should algorithms make layoff decisions? Retrieved from https://www.shrm.org/hr-today/news/hr-magazine/summer-2023/pages/should-algorithms-make-layoff-decisions-.aspx

[286] Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. arXiv:1511.00148. Retrieved from https://arxiv.org/abs/1511.00148

Appendix

A Systematic Review Methodology

This article is an interdisciplinary survey aimed at informing researchers and practitioners interested in fairness and bias in algorithmic hiring. We focused on a Computer Science perspective while summarizing key topics from HR Management, Industrial and Organizational Psychology, Philosophy and Law with mixed methods, including the analysis of influential technical reports, industry white papers and legal literature. Three sections of this work, summarizing measures, mitigation strategies, and data (Sections 4–6), result from a systematic literature review summarized below.

(1) To ensure a broad coverage of the scientific literature centered on computer Science, we leverage three scholarly search engines: IEEE Xplore, ACM Digital Library, and Google Scholar.
(2) We use the query `algorithmic` ∧ `hiring` ∧ `fairness` on article titles, where each term is expanded as follows:
   — `algorithmic`: algorithm* ∨ AI ∨ search* ∨ recommend* ∨ rank* ∨ screen* ∨ retriev*
   — `hiring`: hir* ∨ recruit* ∨ candidate* ∨ job* ∨ work* ∨ resum* ∨ CV ∨ interview* ∨ eval* ∨ appraisal. The last two terms target the evaluation stage.
   — `fairness`: *bias* ∨ *ethic* ∨ *fair* ∨ discriminat* ∨ *equit* ∨ *equal* ∨ *parit* ∨ *symmetr* ∨ gap
(3) To ensure high recall, we consider the top 100 results. To guarantee precision, we manually analyze each article and only select the ones that treat fairness in algorithmic hiring from a quantitative perspective, i.e., performing a fairness audit or introducing a novel method. To exemplify, we discard articles on related yet different topics, such as freelancing [113], focused on human perceptions [159], on qualitative aspects [255], or mitigating biases to improve accuracy without any fairness consideration [49].
(4) We take additional steps to further improve recall. For each included article, we perform forward and backward snowballing [272], pre-filtering article titles with a "`hiring` ∧ `fairness`" query and applying the inclusion criteria in (3). Finally, we consider articles and datasets presented in related surveys [90, 91, 173, 280] or published at dedicated venues,[4] finding one additional dataset (IBM HR Analytics) and no additional article.

Received 21 September 2023; revised 5 July 2024; accepted 30 August 2024

[4]Workshop on Recommender Systems for Human Resources https://recsyshr.aau.dk/