Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices

Manish Raghavan
Cornell University

Solon Barocas
Microsoft Research and Cornell University

Jon Kleinberg
Cornell University

Karen Levy
Cornell University

ABSTRACT

There has been rapidly growing interest in the use of algorithms in hiring, especially as a means to address or mitigate bias. Yet, to date, little is known about how these methods are used in practice. How are algorithmic assessments built, validated, and examined for bias? In this work, we document and analyze the claims and practices of companies offering algorithms for employment assessment. In particular, we identify vendors of algorithmic pre-employment assessments (i.e., algorithms to screen candidates), document what they have disclosed about their development and validation procedures, and evaluate their practices, focusing particularly on efforts to detect and mitigate bias. Our analysis considers both technical and legal perspectives. Technically, we consider the various choices vendors make regarding data collection and prediction targets, and explore the risks and trade-offs that these choices pose. We also discuss how algorithmic de-biasing techniques interface with, and create challenges for, antidiscrimination law.

CCS CONCEPTS

• **Social and professional topics** → **Employment issues**; • **Computing methodologies** → *Machine learning*; • **Applied computing** → *Law*.

KEYWORDS

algorithmic hiring, discrimination law, algorithmic bias

ACM Reference Format:
Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3351095.3372828

1 INTRODUCTION

The study of algorithmic bias and fairness in machine learning has quickly matured into a field of study in its own right, delivering a wide range of formal definitions and quantitative metrics. As industry takes up these tools and accompanying terminology, promises of eliminating algorithmic bias using computational methods have begun to proliferate. In some cases, however, rather than forcing precision and specificity, the existence of formal definitions and metrics has had the paradoxical result of giving undue credence to vague claims about "de-biasing" and "fairness."

In this work, we use algorithmic pre-employment assessment as a case study to show how formal definitions of fairness allow us to ask focused questions about the meaning of "fair" and "unbiased" models. The hiring domain makes for an effective case study because of both its prevalence and its long history of bias. We know from decades of audit studies that employers tend to discriminate against women and ethnic minorities [9, 10, 12, 52], and a recent meta-analysis suggests that little has improved over the past 25 years [75]. Citing evidence that algorithms may help reduce human biases [48, 58], advocates argue for the adoption of algorithmic techniques in hiring [20, 30], with a variety of computational metrics proposed to identify and prevent unfair behavior [35]. But to date, little is known about how these methods are used in practice.

One of the biggest obstacles to empirically characterizing industry practices is the lack of publicly available information. Much technical work has focused on using computational notions of equity and fairness to evaluate specific models or datasets [2, 16]. Indeed, when these models are available, we can and should investigate them to identify potential problems. But what do we do when we have little or no access to models or the data they produce? Certain models may be completely inaccessible to the public, whether for practical or legal reasons, and attempts to audit these models by examining their training data or outputs might jeopardize users' privacy. With algorithmic pre-employment assessments, we find that this is very much the case: models, much less the sensitive employee data used to construct them, are in general kept private. As such, the only information we can consistently glean about industry practices is limited to what companies publicly disclose. Despite this, one of the key findings of our work is that even without access to models or data, we can still learn a considerable amount by investigating what corporations disclose about their practices for developing, validating, and removing bias from these tools.

Documenting claims and evaluating practices. Following a review of firms offering recruitment technologies, we identify 18 vendors of pre-employment assessments. We document what each company has disclosed about its practices and consider the implications of these claims. In so doing, we develop an understanding of industry attempts to mitigate bias and what critical issues are unaddressed.

Prior work has sought to taxonomize the points at which bias can enter machine learning systems, noting that the choice of target variable or outcome to predict, the training data used, and labelling

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02...$15.00
https://doi.org/10.1145/3351095.3372828

of examples are all potential sources of disparities [6, 59]. Following these frameworks, we seek to understand how practitioners handle these key decisions in the machine learning pipeline. In particular, we surface choices and trade-offs vendors face with regard to the collection of data, the ability to validate on representative populations, and the effects of discrimination law on efforts to prevent bias. The heterogeneity we observe in vendors' practices indicates evolving industry norms that are sensitive to concerns of bias but lack clear guidance on how to respond to these worries.

Of course, analyzing publicly available information has its limitations. We are unable, for example, to identify issues that any particular model might raise in practice. Nor can we be sure that vendors aren't doing more behind the scenes to ensure that their models are non-discriminatory. And while other publicly accessible information (e.g., news articles and videos from conferences) might offer further details about vendors' practices, for the sake of consistent comparison, we limit ourselves to statements on vendors' websites. As such, our analysis should not be viewed as exhaustive; however, as we will see, it is still possible to draw meaningful conclusions and characterize industry trends through our methods. One notable limitation we encounter is the lack of information about the validity of these assessments. It is of paramount importance to know the extent to which these tools actually work, but we cannot do so without additional transparency from vendors.

We stress that our analysis is not intended as an exposé of industry practices. Many of the vendors we study exist precisely because they seek to provide a fairer alternative to traditional hiring practices. Our hope is that this work will paint a realistic picture of the landscape of algorithmic techniques in pre-employment assessment and offer recommendations for their effective and appropriate use.

Organization of the rest of the paper. Section 2 contains an overview of pre-employment assessments, their history, and relevant legal precedents. In Section 3, we systematically review vendors of algorithmic screening tools and empirically characterize their practices based on the claims that they make. We analyze these practices in detail in Sections 4 and 5 from technical and legal perspectives, examining ambiguities and particular causes for concern. We provide concluding thoughts and recommendations in Section 6.

2 BACKGROUND

Pre-employment assessments in the hiring pipeline. Hiring decisions are among the most consequential that individuals face, determining key aspects of their lives, including where they live and how much they earn. These decisions are similarly impactful for employers, who face significant financial pressure to make high-quality hires quickly and efficiently [66]. As a result, many employers seek tools with which to optimize their hiring processes.

Broadly speaking, there are four distinct stages of the hiring pipeline, though the boundaries between them are not always rigid: sourcing, screening, interviewing, and selection [14]. Sourcing consists of building a candidate pool, which is then screened to choose a subset to interview. Finally, after candidates are interviewed, selected candidates receive offers. We will focus on *screening*, and in particular, pre-employment assessments that algorithmically evaluate candidates. This includes, for example, questionnaires and video interviews that are analyzed automatically.

Prior work has considered the rise of algorithmic tools in the context of hiring, highlighting the concerns that they raise for fairness. Bogen and Rieke provide an overview of the various ways in which algorithms are being introduced into this pipeline, with a focus on their implications for equity [14]. Garr surveys a number of platforms designed to promote diversity and inclusion in hiring [39]. Sánchez-Monedero et al. [84] analyze some of the vendors considered here from the perspective of UK law, addressing concerns over both discrimination and data protection. Broadly considering the use of data science in HR-related activities, Cappelli et al. identify several practical challenges to the use of algorithmic systems in hiring, and propose a framework to help address them [17]. Ajunwa provides a legal framework to consider the problems algorithmic tools introduce and argues against subjective targets like "cultural fit" [1]. Kim also raises legal concerns over the use of algorithms in hiring in both advertising and screening contexts [56, 57].

Scholars in the field of Industrial-Organizational (IO) Psychology have also begun to grapple with the variety of new pre-employment assessment methods and sources of information enabled by algorithms and big data [44]. Chamorro-Prezumic et al. find that academic research has been unable to keep pace with rapidly evolving technology, allowing vendors to push the boundaries of assessments without rigorous independent research [19]. A 2013 report by the National Research Council summarizes a number of ethical issues that arise in pre-employment assessment, including the role of human intervention, the provision of feedback to candidates, and the goal of hiring for "fit," especially in light of modern data sources [29]. And although proponents argue that pre-employment assessments can push back against human biases [20], assessments (especially data-driven algorithmic ones) run the risk of codifying inequalities while providing a veneer of objectivity.

A history of equity concerns in assessment. Pre-employment assessments date back to examinations for the Chinese civil service thousands of years ago [45]. In the early 1900's, the idea that assessments could reveal innate cognitive abilities gained traction in Western industrial and academic circles, leading to the formation of Industrial Psychology as an academic discipline [40, 53, 68]. During the two World Wars, the U.S. government turned to these assessments in an attempt to quantify soldiers' abilities, paving the way for their widespread adoption in postwar industry [5, 32, 33]. Historically, these assessments were primarily behavioral or cognitive in nature, like the Stanford-Binet IQ test [89], the Myers-Briggs type indicator [69], and the Big Five personality traits [71]. IO Psychology remains a prominent component of these modern assessment tools—many vendors we examine employ IO psychologists who work with data scientists to create and validate assessments.

Cognitive assessments have imposed adverse impacts on minority populations since their introduction into mainstream use [28, 82, 90]. Critics have long contended that observed group differences in test outcomes indicated flaws in the tests themselves [31], and a growing consensus has formed around the idea that while assessments do have some predictive validity, they often disadvantage minorities despite the fact that minority candidates have similar real-world job performance to their white counterparts [28].[1]

[1]Disparities in assessment outcomes for minority populations are not limited to pre-employment assessments. In the education literature, the adverse impact of assessments

The American Psychological Association (APA) recognizes these concerns as examples of "predictive bias" (when an assessment systematically over- or under-predicts scores for a particular group) in its Principles for the Validation and Use of Personnel Selection Procedures [36]. The APA Principles consider several potential definitions of fairness, and while they encourage practitioners to identify and mitigate predictive bias, they explicitly reject the view that fairness requires equal outcomes [36]. As we will see, this focus on predictive bias over outcome-based definitions of fairness forms interesting connections and contrasts with U.S. employment discrimination law.

A brief overview of U.S. employment discrimination law. Title VII of the Civil Rights Act of 1964 forms the basis of regulatory oversight regarding discrimination in employment. It prohibits discrimination with respect to a number of protected attributes ("race, color, religion, sex and national origin"), establishing the Equal Employment Opportunity Commission (EEOC) to ensure compliance [24]. The EEOC, in turn, issued the Uniform Guidelines on Employment Selection Procedures in 1978 to set standards for how employers can choose their employees.

According to the Uniform Guidelines [23], the gold standard for pre-employment assessments is *validity*: the outcome of a test should say something meaningful about a candidate's potential as an employee. The EEOC accepts three forms of evidence for validity: criterion, content, and construct. Criterion validity refers to predictive ability: do test scores correlate with meaningful job outcomes (e.g., sales numbers)? An assessment with content validity tests candidates in similar situations to ones that they will encounter on the job (e.g., a coding interview). Finally, assessments demonstrate construct validity if they test for some fundamental characteristic (e.g., grit or leadership) required for good job performance.

When is an assessment legally considered discriminatory? Based on existing precedent, the Uniform Guidelines provide two avenues to challenge an assessment: disparate treatment and disparate impact [6]. Disparate treatment is relatively straightforward—it is illegal to explicitly treat candidates differently based on categories protected under Title VII [23, 24]. Disparate impact is more nuanced, and while we provide an overview of the process here, we refer the reader to [6] for a more complete discussion.

Under the Uniform Guidelines, the rule of thumb to decide when a disparate impact case can be brought against an employer is the "4/5 rule": if the selection rate for one protected group is less than 4/5 of that of the group with the highest selection rate, the employer may be at risk [23]. If a significant disparity in selection rates is established, an employer may defend itself by showing that its selection procedures are both valid and necessary from a business perspective [23]. Even when a business necessity has been established, an employer can be held liable if the plaintiff can produce an alternative selection procedure with less adverse impact that the employer could have used instead with little business cost [23].[2] Ultimately, both the APA Principles and the Uniform Guidelines agree that validity is fundamental to a good assessment.[3] And while validity can be used as a defense against disparate selection rates, we will see that the Uniform Guidelines' emphasis on outcome disparities and the 4/5 rule significantly impacts vendors' practices.

3 EMPIRICAL FINDINGS

3.1 Methodology

Identifying companies offering algorithmic pre-employment assessments. In order to get a broad overview of the emerging industry surrounding algorithmic pre-employment assessments, we conducted a systematic review of assessment vendors with English-language websites. To identify relevant companies, we consulted Crunchbase's list of the top 300 start-ups (by funding amount) under its "recruiting" category.[4] Crunchbase offers information on public and private companies, providing details on funding and other investment activity. While Crunchbase is not an exhaustive list of all companies working in an industry, it is an often-used resource for tracking developments in start-up companies. Companies can create profiles for themselves, subject to validation.[5] We supplemented this list with an inventory of relevant companies found in recent reports by Upturn [14], a technology research and advocacy firm focused on civil rights, and RedThread Research [39], a research and advisory firm specializing in new technologies for human resource management. This resulted in 22 additional companies, for a combined total of 322. There was substantial overlap between the three sources considered.

Thirty-nine of these companies did not have English-language websites, so we excluded them. Recall that the hiring pipeline has four primary stages (sourcing, screening, interviewing, and selection); we ruled out vendors that do not provide assessment services at the screening stage, leaving us with 45 vendors. Note that this excluded companies that merely provide online job boards or marketplaces like Monster.com and Upwork. Twenty-two of the remaining vendors did not obviously use any predictive technology (e.g., coding interview platforms that only evaluated correctness or rule-based screening) or did not offer explicit assessments (e.g., scraping candidate information from other sources), and an additional 5 did not provide enough information for us to make concrete determinations, leaving us with 18 vendors in our sample. With these 18 vendors, in April 2019,[6] we recorded administrative information available on Crunchbase (approximate number of employees, location, and total funding) and undertook a review of their claims and practices, which we explain below.

Documenting vendors' claims and practices. Based on prior frameworks intended to interrogate machine learning pipelines for bias [6, 59], we ask the following questions of vendors:

on minorities is well-documented [64]. This has led to a decades-long line of literature seeking to measure and mitigate the observed disparities (see [50] for a survey).

[2]It should be noted that this description is based on a particular (although the most common) interpretation of Title VII. Some legal scholars contend that Title VII offers stronger protections to minorities [15, 54], and there is disagreement on how (or whether) to operationalize the 4/5 rule through statistical tests [21, 22, 86, 87]. In this work, we will not consider alternative interpretations of Title VII, nor will we get into the specifics of how exactly to detect violations of the 4/5 rule.

[3]Many psychologists disagree with the specific conception of validity endorsed by the Uniform Guidelines [13, 67, 83]; however, there is broad agreement that some form of validation is necessary.

[4]https://www.crunchbase.com/hub/recruiting-startups

[5]https://support.crunchbase.com/hc/en-us/articles/115011823988-Create-a-Crunchbase-Profile

[6]Our empirical findings are specific to this moment in time; practices and documentation may have changed since then.

- What types of assessments do they provide (e.g., questions, video interviews, or games)? [Features]
- What is the outcome or quality that these assessments aim to predict (e.g., sales revenue, annual review score, or grit)? [Target variable]
- What data are used to develop the assessment (e.g., the client's or the vendor's own data)? [Training data]
- What information do they provide regarding validation processes (e.g., validation studies or whitepapers)? [Validation]
- What claims or guarantees (if any) are made regarding bias or fairness? When applicable, how do they achieve these guarantees? [Fairness]

To answer these questions, we exhaustively searched the websites of each company. This included following all internal links, downloading any reports or whitepapers they provided, and watching webinars found on their websites. Almost all vendors provided an option to request a demo; we avoided doing so since our focus is on accessible and public information. Sometimes, company websites were quite sparse on information, and we were unable to conclusively answer all questions for all vendors.

3.2 Findings

In our review, we found 18 vendors providing algorithmically driven pre-employment assessments. Those that had available funding information on Crunchbase (16 out of 18) ranged in funding from around $1 million to $93 million. Most vendors (14) had 50 or fewer employees, and half (9) were based in the United States. 15 vendors were present in Crunchbase's "Recruiting Startups" list; the remaining vendors were taken from reports by Upturn [14] and RedThread Research [39]. Many vendors were present in all of these sources. Table 1 summarizes our findings. Table 3 in Appendix A contains administrative information about the vendors we included.

Assessment types. The types of assessments offered varied by vendor. The most popular assessment types were questions (11 vendors), video interview analysis (6 vendors), and gameplay (e.g., puzzles or video games) (6 vendors). Note that many vendors offered multiple types of assessments. Question-based assessments included personality tests, situational judgment tests, and other formats. For video interviews, candidates were typically either asked to record answers to particular questions or more free-form "video resumes" highlighting their strengths. These videos are then algorithmically analyzed by vendors.

Target variables and training data. Most of the vendors (15) offer custom or customizable assessments, adapting the assessment to the client's particular data or job requirements. In practice, decisions about target variables and training data are made together based on where the data come from. Eight vendors build assessments based on data from the client's past and current employees (see Figure 1). Vendors in general leave it up to clients to determine what outcomes they want to predict, including, for example, performance reviews, sales numbers, and retention time. Other vendors who offer customizable assessments without using client data either use human expertise to determine which of a pre-determined set of competencies are most relevant to the particular job (the vendor's analysis of a job role or a client's knowledge of relevant requirements) or don't explicitly specify their prediction targets. In such cases, the vendor provides an assessment that scores applicants on various competencies, which are then combined into a "fit" score based on a custom formula. Thus, even among vendors who tailor their assessments to a client, they do so in different ways.

Vendors who only offer pre-built assessments typically either provide assessments designed for a particular job role (e.g., salesperson), or provide a sort of "competency report" with scores on a number of cognitive or behavioral traits (e.g., leadership, grit, teamwork). These assessments are closer in spirit to traditional psychometric assessments like the Myers-Briggs Type Indicator or Big Five Personality Test; however, unlike traditional assessments that rely on a small number of questions, modern assessments may build psychographic profiles using machine learning to analyze rich data sources like a video interview or gameplay.

Validation. Generally, vendors' websites do not make clear whether vendors validate their models, what validation methodologies they use, how they select validation data, or how validation procedures might be tailored to the particular client. Good & Co.,[7] notably, provides fairly rigorous validation studies of the psychometric component of their assessment, as well as a detailed audit of how the scores differ across demographic groups; however, they do not provide similar documentation justifying the algorithmic techniques they use to recommend candidates based on "culture fit."

Accounting for bias. In total, while 15 of the vendors made at least abstract references to "bias" (sometimes in the context of well-established human bias in hiring), only 7 vendors explicitly discussed compliance or adverse impact with respect to the assessments they offered. Three vendors explicitly mentioned the 4/5 rule, and an additional 4 advertised "compliance" or claimed to control adverse impact more generally. Several of these vendors claimed to test models for bias, "fixing" it when it appeared. HireVue and pymetrics, in particular, offered a detailed description of their overall approaches to de-biasing, which involves removing features correlated with protected attributes when adverse impact is detected. Other vendors (e.g., Knockri and PredictiveHire) claimed to "fix" adverse impact when it is found without going into further detail.

Among those that do make concrete claims, all vendors we examined specifically focus on equality of outcomes and compliance with the 4/5 rule. Roughly speaking, there are two ways in which vendors claim to achieve these goals: naturally unbiased assessments and active algorithmic de-biasing. Typically, vendors claiming to provide naturally unbiased assessments seek to measure underlying cognitive or behavioral traits, so their assessments output a small number of scores, one for each competency being measured. In this setting, a naturally unbiased assessment in one that produces similar score distributions across demographic groups. Koru, for instance, measures 7 traits (e.g., "grit" and "presence") and claims that "[i]n all panels since 2015, the Pre-Hire assessment does not show bias against women or minority respondents" [51].

Other vendors actively intervene in their learned models to remove biases. One technique that we have observed across multiple vendors (e.g., HireVue, pymetrics, PredictiveHire) is the following:

[7] https://good.co/

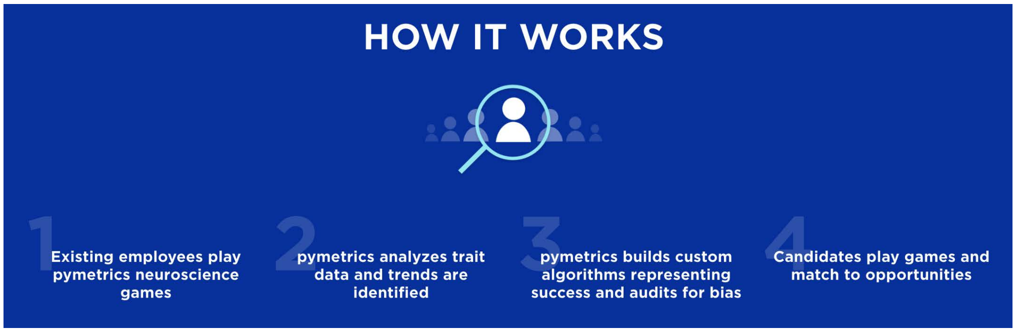

Figure 1: Description of the pymetrics process (screenshot from the pymetrics website: https://www.pymetrics.com/employers/)

Vendor name	Assessment types [Features]	Custom? [Target & Training data]	Validation info [Validation]	Adverse impact [Fairness]
8 and Above	phone, video	S	–	bias mentioned
ActiView	VR assessment	C	validation claimed	bias mentioned
Assessment Innovation	games, questions	–	–	bias mentioned
Good&Co	questions	C, P	multiple studies	adverse impact
Harver	games, questions	S	–	–
HireVue	games, questions, video	C, P	–	4/5 rule
impress.ai	questions	S	–	–
Knockri	video	S	–	bias mentioned
Koru	questions	S	some description	adverse impact
LaunchPad Recruits	questions, video	–	–	bias mentioned
myInterview	video	–	–	compliance
Plum.io	questions, games	S	validation claimed	bias mentioned
PredictiveHire	questions	C	–	4/5 rule
pymetrics	games	C	small case study	4/5 rule
Scoutible	games	C	–	–
Teamscope	questions	S, P	–	bias mentioned
ThriveMap	questions	C	–	bias mentioned
Yobs	video	C, S	–	adverse impact

Table 1: Examining the websites of vendors of algorithmic pre-employment assessments, we answer a number of questions regarding their assessments in relation to questions of fairness and bias. This involves exhaustively searching their websites, downloading whitepapers they provide, and watching webinars they make available. This table presents our findings. The "Assessment types" column gives the types of assessments each vendor offers. In the "Custom?" column, we consider the source of data used to build an assessment: C denotes "custom" (uses employer data), S denotes "semi-custom" (qualitatively tailored to employer without data) and P denotes "pre-built." The "Validation?" column contains information vendors publicly provided about their validation processes. In the "Adverse impact" column, we recorded phrases found on vendors' websites addressing concerns over bias.

build a model and test it for adverse impact against various subgroups.[8] As Bogen and Rieke also observe [14], if adverse impact is found, the model and/or data are modified to try to remove it, and then the model is tested again for adverse impact. HireVue and pymetrics downweight or remove features found to be highly correlated with the protected attribute in question, noting that this can significantly reduce adverse impact with little effect on the predictive accuracy of the assessment. This is done prior to the model's deployment on actual applicants, though some vendors claim to periodically test and update models. In Section 5, we discuss in depth these efforts to define and remove bias.

[8]pymetrics, for instance, open-sources the tests it uses: https://github.com/pymetrics/audit-ai

4 ANALYSIS OF TECHNICAL CONCERNS

Our findings in Section 3 raise several technical challenges for the pre-employment assessment process. In this section, we focus on two areas that are particularly salient in the context of algorithmic hiring: **data choices**, where vendors must decide where to draw data from and what outcomes to predict; and the use of **alternative**

Vendor	Claim about bias
HireVue	Provide "a highly valid, bias-mitigated assessment"
pymetrics	"…the Pre-Hire assessment does not show bias against women or minority respondents."
PredictiveHire	"AI bias is testable, hence fixable."
Knockri	"Knockri's A.I. is unbiased because of its full spectrum database that ensures there's no benchmark of what the 'ideal candidate' looks like."

Table 2: Examples of claims that vendors make about bias, taken from their websites.

assessment formats, like game- or video-based assessments that rely on larger feature sets and more complex machine learning tools than traditional question-based assessments.

4.1 Data Choices

Machine learning is often viewed as a process by which we predict a given output from a given set of inputs. In reality, neither the inputs nor outputs are fixed. Where do the data come from? What is the "right" outcome to predict? These and others are crucial decisions in the machine learning pipeline, and can create opportunities for bias to enter the process.

Custom assessments. Consider a hypothetical practitioner building a custom assessment to identify the "best" candidates for her client. As is the case in many domains, translating this to a feasible data-driven task forces our practitioner to make certain compromises [72]. It quickly becomes clear that she must operationalize "best" in some measurable way. What does the client value? Sales numbers? Cultural fit? Retention? And, crucially, what data does the client have? This is a nontrivial constraint: many companies don't maintain comprehensive and accessible data about employee performance, and thus, a practitioner may be forced to do the best she can with the limited data that she is given [17]. Note that relying on the client's data has already forced the practitioner to only learn from the client's existing employees; at the outset, at least, she has data on how those who *weren't* hired would have performed.

Once a target is identified, the practitioner needs a dataset on which to train a model. Since she has performance data on previous employees, she needs them to take the assessment so she can link their scores to their observed job performance. How many employees' data does she need in order to get an accurate model? What if certain employees don't want to or don't have time to take the assessment? Is the set of employees who respond representative of the larger applicant pool who will ultimately be assessed?

Finally, the practitioner is in a position to actually build a model. Along the way, however, she had to make several key choices, often based on factors (like client data availability) outside her control. The choice of target variable is particularly salient. Proxies like job evaluations, for instance, can exhibit biases against minorities [70, 79, 88]. Moreover, predicting the success of future employees based on current employees inherently skews the task toward finding candidates resembling those who have already been hired.

Some vendors go beyond trying to identify candidates who are generically good, or even good for a particular client, and explicitly focus on finding candidates who "fit" with an existing employee or team. Both Good & Co. and Teamscope provide these team-specific tools for employers, and Good & Co. further advertises their assessments as a way to "[r]eplicate your top performers."[9] If models are localized to predict fit with particular teams, any role at any company could in principle have its own tailor-made predictive model. But when models are customized at such a small scale, it can be quite difficult to determine what it means for such a model to be biased or discriminatory. Does each team-specific model need to be audited for bias? How would a vendor go about doing so?

And yet, while it is easy to criticize vendors for their choices, it's not clear that there are better alternatives. In practice, it is impossible to even define, let alone collect data on, an objective measure of a "good" employee. Nor is it always feasible to get completely representative data. Vendors and advocates point out that many of the potentially problematic elements here (subjective evaluations; biased historical samples; emphasis on fit) are equally present, if not more so, in traditional human hiring practices [20].

Customizable and pre-built assessments. Instead of building a new custom assessment for each client, it may be tempting to instead offer a pre-built assessment (perhaps specific to a particular job role) that has been validated across data from a variety of clients. This approach has its advantages: it isn't subject to the idiosyncratic data of each client, and it can draw from a diverse range of candidates and employees to learn a broad notion of what a "good" employee looks like. Additionally, pre-built assessments may be attractive to clients with too few existing employees to build a custom assessment.

Some vendors offer assessments that are mostly pre-built but somewhat customizable. Koru and Plum.io, for example, provide pre-built assessments to evaluate a fixed number of competencies. Experts then analyze the job description and role for a particular client and determine which competencies are most important for the client's needs. Thus, these vendors hope to get the best of both worlds: assessments validated on large populations that are still flexible enough to adapt to the specific requirements of each client. As shown in Figure 2, the firm 8 and Above profiles over 60 traits based on a video interview, but also reports a single "Elev8" score tailored to the particular client.

Despite these benefits, pre-built assessments do have drawbacks. Individual competencies like "grit" or "openness" are themselves constructs, and attempts to measure them must rely on other psychometric assessments as "ground truth." Given that traits can be measured by multiple tests that don't perfectly correlate with one another [81], it may be difficult to create an objective benchmark

[9] https://good.co/pro/

against which to compare an algorithmic assessment. Furthermore, it is generally considered good practice to build and validate assessments on a representative population for a particular job role [36], and both underlying candidate pools and job specifics differ across locations, companies, and job descriptions. Pre-built assessments must by nature be general, but as a consequence, they may not adapt well to the client's requirements.

Necessary trade-offs. This leads to an inherently challenging technical problem: on the one hand, more data is usually beneficial in creating and validating an assessment; on the other hand, drawing upon data from related but somewhat different sources may lead to inaccurate conclusions. We can view this as an instance of domain adaptation and the bias-variance tradeoff, well studied in the statistics and machine learning literature [8, 37]. Pooling data from multiple companies or geographic locations may reduce variance due to small sample sizes at a particular company, but comes at the cost of biasing the outcomes away from the client's specific needs. There is no obvious answer or clear best practice here, and vendors and clients must carefully consider the pros and cons of various assessment types. Larger clients may be better positioned for vendors to build custom assessments based solely on their data; smaller clients may turn to pre-built assessments, making the assumption that the candidate pool and job role on which the assessment was built is sufficiently similar to warrant generalizing its conclusions.

4.2 Alternative Assessment Formats

Once an assessment has been built, it must be validated to verify that it performs as expected. Psychologists have developed extensive standards to guide assessment creators in this process [36]; however, modern assessment vendors are pushing the boundaries of assessment formats far beyond the pen-and-paper tests of old, often with little regulatory oversight [19]. Game- and video-based assessments, in particular, are increasingly common. Vendors point to an emerging line of literature showing that features derived from these modern assessment formats correlate with job outcomes and personality traits [42, 60] as evidence that these assessments contain information that can be predictive of job outcomes, though they rarely release rigorous validation studies of their own.

Technical challenges for alternative assessments. While there is evidence for the predictive validity of alternative assessments, empirical correlation is no substitute for theoretical justification. Historically, IO psychologists have designed assessments based on their research-driven knowledge that certain traits correlate with desirable outcomes. To some extent, machine learning attempts to automate this process by discovering relationships (e.g., between actions in a video game and personality traits) instead of quantifying known relationships. Of course, machine learning can be used to unearth meaningful relationships. But it may also find relationships that experts don't understand. When the expert is unable to explain why, for example, the cadence of a candidate's voice indicates higher job performance, or why reaction time predicts employee retention, should a vendor rely on these features? From a technical perspective, correlations that cannot be theoretically justified may fail to generalize well or remain stable over time, and, in light of such concerns, the APA Principles caution that a practitioner should "establish a clear rationale for linking the resulting scores to the criterion constructs of interest" [36]. Yet when an algorithm takes in "millions of data points" for each candidate (as advertised by pymetrics[10]), it may not be possible to provide a qualitative justification for the inclusion of each feature.

Moreover, automated discovery of relationships makes it difficult for a critical expert to detect when the model makes indirect use of a proscribed characteristic. Rich sources of data can easily encode properties that are illegal to use in the hiring process. Facial analysis, in particular, has been heavily scrutinized recently. A wave of studies has shown that several commercially available facial analysis techniques suffer from disparities in error rates across gender and racial lines [16, 76, 78], and more broadly, evidence suggests that we may not be able to reliably infer emotions from facial expressions, especially cross-culturally [7]. Concerns have also been raised over the use of affect and emotion recognition for those with disabilities, particularly in the context of employment [38, 43, 49].

Because it can be quite expensive and technically challenging to build facial analysis software in-house, vendors will often turn to third parties (e.g., Affectiva[11]) who provide facial analysis as a service. As a result, vendors lack the ability or resources to thoroughly audit the software they use. With these concerns in mind, U.S. Senators Kamala Harris, Patty Murray, and Elizabeth Warren recently wrote a letter to the EEOC asking for a report on the legality and potential issues with the use of facial analysis in pre-employment assessments [46]. Even more recently, Illinois passed a law requiring applicants to be notified and provide consent if their video interviews will be analyzed by artificial intelligence [3], though it's not clear what happens if an applicant refuses to consent.

While heightened publicity regarding racial disparities in facial analysis has prompted many third-party vendors of this technology to respond by improving the performance of their tools on minority populations [74, 80], it remains unclear what information facial analysis relies on to draw conclusions about candidates. Facial expressions may contain information about a range of sensitive attributes from obvious ones like ethnicity, gender, and age to more subtle traits like a candidate's mental and physical health [60, 96].[12] Given the opacity of the deep learning models used for facial analysis, it can be difficult or even impossible to detect if a model inadvertently learns proxies for prohibited features.

5 ALGORITHMIC DE-BIASING

Under Title VII, employers bear ultimate legal responsibility for their hiring decisions. Employers, then, remain strongly motivated to mitigate their potential liability against disparate impact claims. Vendors, in turn, are incentivized to build demonstrably unbiased tools that help employers to avoid such liability.

As we have described, all vendors in our sample who made concrete claims about de-biasing (including the two best-funded firms in our sample) did so with reference to equality of outcomes and compliance with the 4⁄5 rule. In this section, we explore the effects of this reliance on the stages of a typical disparate impact lawsuit. We then explore technical approaches that have been proposed

[10]https://perma.cc/3284-WTS8
[11]https://www.affectiva.com/
[12]As a general matter, the Americans with Disabilities Act prohibits employers from collecting or considering information about candidates' health [25].

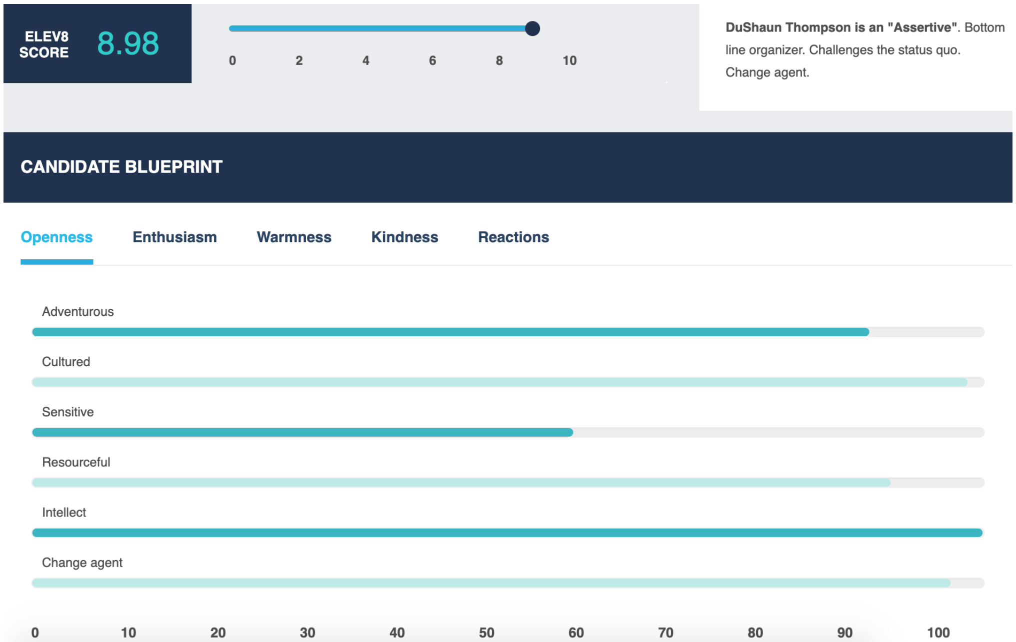

Figure 2: Part of a sample candidate profile from 8 and Above, based on a 30-second recorded video cover letter (screenshot from the 8 and Above website: https://www.8andabove.com/p/profile/blueprint/643)

to control outcome disparities, and their relationship to the law. Finally, we describe some important consequences of the de-biasing strategies favored by vendors.

5.1 Algorithmic De-Biasing and Disparate Impact Litigation

Recall the three steps in a disparate impact case. The plaintiff must first establish that the employer's selection procedure generates a disparate impact. Once established, the employer must then defend itself by justifying the disparate impact by reference to some business necessity. In this case, an employer would likely do so by establishing the validity of the model driving its hiring decisions. Finally, the plaintiff may then challenge the proffered justification as faulty or demonstrate that an alternative practice exists that would serve the employer's business objective equally well while reducing the disparate impact in its selection rates.

Note that disparate impact doctrine does not prohibit disparate impact altogether; it renders employers liable for an *unjustified* or *avoidable* disparate impact. Vendors' choice to enforce the 4/5 rule might therefore seem overly cautious: although employers could justify an assessment that has a disparate impact by demonstrating its validity (as we discuss in Section 2), vendors take steps to ensure that employers are not placed in this position, because assessments are prevented from having a disparate impact in the first place. One possible explanation for adopting the 4/5 rule is that vendors might be catering to employers' aversion to legal risk.

As to the second step, the practical effect of vendors' reliance on the 4/5 rule is to obviate the need for an employer to demonstrate business necessity through a legally rigorous validation process. According to the Uniform Guidelines, employers only need to validate their selection procedure if it has a disparate impact. Of course, clients might still expect and even demand validation studies from vendors, given their goal of selecting qualified candidates. As a consequence, the choice of how to validate seems to become a *business* decision rather than a *legal* imperative.

The final step in a disparate impact case raises yet another possible explanation for vendors' decisions to adopt the 4/5 rule as a constraint. Recall that, at this stage, employers bear liability if they failed to adopt an alternative practice that could have minimized any business-justified disparity created by their selection procedure, provided that such practices were not too costly. Employers therefore run significant legal risks if they do not take such steps. In turn, should vendors have some way to minimize disparity without sacrificing the accuracy of their assessments, failing to do so might place their clients in legal jeopardy. A plaintiff could assert that this very possibility reveals that any evident disparate impact—even if justified by a validation study—was avoidable.

While the burden of identifying this alternative business practice rests with the plaintiff, vendors may want to preempt this argument by taking affirmative steps to explore how to minimize disparate impact without imposing unwelcome costs on the employer. In the past, such exploratory efforts might have been costly and difficult, since discovering an alternative business practice that is equally effective for the firm, while generating less disparity in selection rates, was no easy task. Many modern assessments (e.g., those with a large number of features) make some degree of exploration almost trivial, allowing vendors to find a model that (nearly) maintains maximum accuracy while reducing disparate impact.

In this way, the ready availability of algorithmic techniques might effectively create a legal *imperative* to use them. If the adverse impact of a business-justified model could be reduced through algorithmic de-biasing—without significantly harming predictive ability, and at trivial cost—de-biasing itself might be considered an "alternative business practice," and therefore render the employer liable for not adopting it.

5.2 Methods to Control Outcome Disparities

Thus, for legal reasons, a vendor may choose to control outcome disparities in strict adherence to the 4⁄5 rule. But this is not the end of the story; multiple techniques exist to control outcome differences. Here, we explore both historical and contemporary approaches in comparison with the de-biasing techniques we observe.

The most straightforward approach to control outcome differences is known as "within-group scoring," under which scores are reported as a percentile with respect to the particular group in question. Employers could then select candidates above a particular threshold for each group (top 10% from Group A, top 10% from Group B, etc.), which would naturally result in equal selection rates. Recall that in the de-biasing reviewed above, vendors achieve (approximately) equal selection rates by systematically removing features from the model that contribute to a disparate impact. In so doing, they may lose useful information contained in these features as well, undermining their ability to maintain an accurate rank order within each group. In contrast, within-group scoring may theoretically be the optimal way to equalize selection rates, since it preserves rank order [27, 63].

In fact, within-group scoring was used for the General Aptitude Test Battery (GATB), a pre-employment assessment developed in the 1940s by the US Employment Service (USES), due to significant differences in score distributions across ethnic groups. In particular, the USES reported results as within-group percentile scores by ethnicity—black, Hispanic, and other [28, 85]. Commissioned to investigate the justification for such a policy, a National Academy of Sciences study recommended the continued use of within-group percentiles because without them, minority applicants would suffer from "higher false-rejection rates" [28].

In principle, within-group score reporting (also known as "race-norming") would satisfy the 4⁄5 rule; so why don't vendors use it? In fact, within-group reporting would likely be considered illegal today. In 1986 the Department of Justice challenged its legality in the GATB, claiming that it constituted disparate treatment [85], and the practice was prohibited by the Civil Rights Act of 1991 [26].

This points to a longstanding tension between disparate treatment and disparate impact: some techniques to control outcome disparities require the use of protected attributes, which may be considered disparate treatment. To circumvent this, the vendors we observe engaging in algorithmic de-biasing take into account protected attributes when *building* models, but ultimately produce models that do not take protected attributes as input. In this way, individual decisions do not exhibit disparate treatment, and yet, outcome disparities can still be mitigated.

In fact, these techniques fit into a broader category of methods known as Disparate Learning Processes (DLPs), a family of algorithms designed to produce decision rules that (approximately) equalize outcomes without engaging in disparate treatment at the individual level [63, 73, 93]. There are slight differences between DLPs as found in the computer science literature and vendors' algorithmic de-biasing efforts: DLPs typically work by imposing constraints that prevent outcome disparities on the learning algorithm that produces the model; the algorithmic de-biasing we observe, on the other hand, simply removes features correlated with protected attributes until outcomes are within a tolerable range. In spirit, however, these techniques are ultimately quite related.

Similar connections exist to "fair representation" learning, where an "encoder" is built to process data by removing information about protected attributes, including proxies and correlations [34, 65, 95]. Thus, any model built on data processed by the encoder would have approximately equal outcomes, since outputs of the encoder contain very little information about protected attributes. As in DLPs, protected attributes are used only to create the encoder; after deployment, when the encoder processes any individual's data, it does not have access to protected attributes. We can think of some vendors' practices as analogous to building such an encoder—one that "processes" data by simply discarding features highly correlated with protected attributes.

5.3 Limitations of Outcome-Based De-Biasing

Despite the perhaps good reasons vendors have to use the particular form of algorithmic de-biasing discussed above, these techniques face important caveats and consequences worth mentioning.

Outcome-based notions of bias are intimately tied to the datasets on which they are evaluated. As both the EEOC Guidelines and APA Principles clearly articulate, a representative sample is crucial for validation [23, 36]. The same holds true for claims regarding outcome disparities: they may depend on whether the assessment is taken by recent college grads in Michigan applying for sales positions or high school dropouts in New York applying for jobs stocking warehouses. Thus, when evaluating claims regarding outcome disparities, it is critical to understand how vendors collect and maintain relevant, representative data.

While outcome disparities are important for vendors to consider, especially in light of U.S. regulations, discrimination and the 4⁄5 rule should not be conflated. Vendors may find it necessary from a legal or business perspective to build models that satisfy the 4⁄5 rule, but this is not a substitute for a critical analysis into the mechanisms by which bias and harm manifest in an assessment. For example, differential validity, which occurs when an assessment is better at ranking members of one group than another, should be a top-level concern when examining an assessment [36, 92]. But because of the legal emphasis placed on adverse impact, vendors have little incentive to structure their techniques around it. Furthermore, it can be challenging to identify and mitigate outcome disparities with respect to protected attributes employers typically don't collect (e.g., sexual orientation [62]). In such cases, vendors may need to consider alternative approaches to prevent discrimination.

More broadly, bias is not limited to the task of predicting outputs from inputs; a critical, holistic examination of the entire assessment development pipeline may surface deeper concerns. Where do inputs and outputs come from, and what justification do they have? Are there features that shouldn't be used? This isn't to say that some vendors are not already asking these questions; however, in the interest of forming industry standards surrounding algorithmic assessments, the legal operationalization of the 4/5 rule as a definition of bias runs the risk of downplaying the importance of examining a system as a whole.

Both the law and existing techniques focus on assessment outcomes as binary (screened in/out); however, some platforms actually rank candidates (explicitly, or implicitly by assigning numerical scores). While screening decisions can ultimately be viewed as binary (a candidate is either interviewed or not), there are a number of subtleties induced by ranking: a lower-ranked candidate may only be interviewed after higher-ranked candidates, and their lower score could unduly bias future decision-makers against them [14]. There is no clear analog of the 4/5 rule for ranking; in practice, vendors may choose a cut-off score and test for adverse impact via the 4/5 rule [4, 23]. In the computer science literature, there are ongoing efforts to define technical constraints on rankings in the spirit of equal representation and the 4/5 rule [18, 91, 94], and LinkedIn has adopted a similar approach to encourage demographic diversity in its search results [41]. However, none of these approaches has received any sort of consensus or official endorsement.

From a policy perspective, the EEOC can and should clarify its position on the use of algorithmic de-biasing techniques, which to our knowledge has yet to be challenged in court. Legal scholars have begun to debate the legality of "algorithmic affirmative action" in various contexts [11, 47, 55, 61, 77], but the debate is far from settled. While existing guidelines can be argued to apply to ML-based assessments, the de-biasing techniques described above do present new opportunities and challenges.

6 DISCUSSION AND RECOMMENDATIONS

In this work, we have presented an in-depth analysis into the bias-related practices of vendors of algorithmic pre-employment assessments. Our findings have implications not only for hiring pipelines, but more broadly for investigations into algorithmic and socio-technical systems. Given the proprietary and sensitive nature of models built for actual clients, it is often infeasible for external researchers to perform a traditional audit; despite this, we are able to glean valuable information by delving into vendors' publicly available statements. Broadly speaking, models result from the application of a **vendor's practices** to a real-world setting. Thus, by learning about these practices, we can draw conclusions and raise relevant questions about the resultant models. In doing so, we can create a common vocabulary with which we can discuss and compare practices. We found it useful to **limit the scope** of our inquiry in order to be able to ask and answer concrete questions. Even just considering algorithms used in the context of hiring, we found enough heterogeneity (as have previous reports on the subject [14, 39]) that it was necessary to further refine our focus to those used in pre-employment assessments. While this did lead us to exclude a number of innovative and intriguing hiring technologies (see, e.g., Textio[13] or Jopwell[14]), it allowed us to make specific and direct comparisons between vendors and get a more detailed understanding of the technical challenges specific to assessments.

In analyzing models via practices, we observe that it is crucial to consider technical systems in conjunction with the **context** surrounding their use and deployment. It would be difficult to understand vendors' design decisions without paying attention to the relevant legal, historical, and social influences. Moreover, in order to push beyond hypothetical or anecdotal accounts of algorithmic bias, we need to incorporate empirical evidence from the field.

Based on our findings, we summarize the following policy recommendations discussed throughout this work.

(1) Transparency is crucial to further our understanding of these systems. While there are some exceptions, vendors in general are not particularly forthcoming about their practices. Additional transparency is necessary to craft effective policy and enable meaningful oversight.
(2) Disparate impact is not the only indicator of bias. Vendors should also monitor other metrics like differential validity.
(3) Outcome-based measures of bias (including tests for disparate impact and differential validity) are limited in their power. They require representative datasets for particular applicant pools and fail to critically examine the appropriateness of individual predictors. Moreover, they depend on access to protected attributes that are not always available.
(4) We may need to reconsider legal standards of validity under the Uniform Guidelines in light of machine learning. Because machine learning may discover relationships that we do not understand, a statistically valid assessment may inadvertently leverage ethically problematic correlations.
(5) Algorithmic de-biasing techniques have significant implications for "alternative business practices," since they automate the search for less discriminatory alternatives. Vendors should explore these techniques to reduce disparate impact, and the EEOC should offer clarity about how the law applies.

Our work leads naturally to a range of questions, ranging from those that seem quite technical (What is the effect of algorithmic de-biasing on model outputs? When should data from other sources be incorporated?) to socio-political (What additional regulatory constraints could improve the use of algorithms in assessment? How can assessments promote the autonomy and dignity of candidates?). Because the systems we examine are shaped by technical, legal, political, and social forces, we believe that an interdisciplinary approach is necessary to get a broader picture of both the problems they face and the potential avenues for improvement.

ACKNOWLEDGMENTS

We thank Rediet Abebe, Ifeoma Ajunwa, Bilan Ali, Lewis Baker, Miranda Bogen, Heather Bussing, Albert Chang, A. F. Cooper, Fernando Delgado, Kate Donahue, Stacia Garr, Avital Gertner-Samet, Jim Guszcza, Stephen Hilgartner, Lauren Kilgour, Loren Larsen, Richard Marr, Cassidy McGovern, Helen Nissenbaum, Samir Passi, David Pedulla, Frida Polli, Sarah Riley, David Robinson, Caleb Rottman, John Sumser, Kelly Trindel, Katie Van Koevering, Briana Vecchione, Suresh Venkatasubramanian, Angela Zhou, Malte Ziewitz, and Lindsey Zuloaga for their suggestions and insights.

[13]Textio (https://textio.com/) analyzes job descriptions for gender bias and makes suggestions for alternative, gender-neutral framings.
[14]Jopwell (https://www.jopwell.com/) builds and maintains a network of Black, Latinx, and Native American students and connects students these with employers.

REFERENCES

[1] Ifeoma Ajunwa. 2020. The Paradox of Automation as Anti-Bias Intervention. *Cardozo Law Review* 41 (2020).

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016).

[3] Illinois General Assembly. 2019. Artificial Intelligence Video Interview Act.

[4] Lewis Baker, David Weisberger, Daniel Diamond, Mark Ward, and Joe Naso. 2018. audit-AI. https://github.com/pymetrics/audit-ai.

[5] Loren Baritz. 1960. *The servants of power: A history of the use of social science in American industry.* Wesleyan University Press.

[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[7] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.

[8] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79, 1-2 (2010), 151–175.

[9] Marc Bendick and Ana P Nunes. 2012. Developing the research basis for controlling bias in hiring. *Journal of Social Issues* 68, 2 (2012), 238–262.

[10] Marc Bendick Jr, Charles W Jackson, and J Horacio Romero. 1997. Employment discrimination against older workers: An experimental study of hiring practices. *Journal of Aging & Social Policy* 8, 4 (1997), 25–46.

[11] Jason R Bent. 2020. Is Algorithmic Affirmative Action Legal? *Georgetown Law Journal* 108 (2020).

[12] 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.

[13] Daniel A Biddle. 2008. Are the uniform guidelines outdated? Federal guidelines, professional standards, and validity generalization (VG). *The Industrial-Organizational Psychologist* 45, 4 (2008), 17–23.

[14] Miranda Bogen and Aaron Rieke. 2018. *Help Wanted: An Exploration of Hiring Algorithms, Equity, and Bias.* Technical Report. Upturn. https://www.upturn.org/static/reports/2018/hiring-algorithms/files/Upturn%20--%20Help%20Wanted%20-%20An%20Exploration%20of%20Hiring%20Algorithms,%20Equity%20and%20Bias.pdf

[15] Stephanie Bornstein. 2018. Antidiscriminatory Algorithms. *Ala. L. Rev.* 70 (2018), 519.

[16] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency.* 77–91.

[17] Peter Cappelli, Prasanna Tambe, and Valery Yakubovich. 2018. Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. *Available at SSRN 3263878* (2018).

[18] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic.* 28:1–28:15. https://doi.org/10.4230/LIPIcs.ICALP.2018.28

[19] Tomas Chamorro-Premuzic, Dave Winsborough, Ryne A Sherman, and Robert Hogan. 2016. New talent signals: Shiny new objects or a brave new world? *Industrial and Organizational Psychology* 9, 3 (2016), 621–640.

[20] Tomas Chamorro-Prezumic and Reece Akhtar. 2019. Should Companies Use AI to Assess Job Candidates? *Harvard Business Review* (2019).

[21] Richard M Cohn. 1979. On the Use of Statistics in Employment Discrimination Cases. *Ind. LJ* 55 (1979), 493.

[22] Richard M Cohn. 1979. Statistical Laws and the Use of Statistics in Law: A Rejoinder to Professor Shoben. *Ind. LJ* 55 (1979), 537.

[23] Equal Employment Opportunity Commission, Civil Service Commission, et al. 1978. Uniform guidelines on employee selection procedures. *Federal Register* 43, 166 (1978), 38290–38315.

[24] U.S. Congress. 1964. Civil Rights Act.

[25] U.S. Congress. 1990. Americans with Disabilities Act.

[26] U.S. Congress. 1991. Civil Rights Act.

[27] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 797–806.

[28] National Research Council et al. 1989. *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* National Academies Press.

[29] National Research Council et al. 2013. *New directions in assessing performance potential of individuals and groups: Workshop summary.* National Academies Press.

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

[31] Hamilton Cravens. 1978. *The triumph of evolution: The heredity–environment controversy, 1900–1941.* Johns Hopkins University Press.

[32] Philip Hunter DuBois. 1970. *A history of psychological testing.* Allyn and Bacon.

[33] Marvin D Dunnette and Walter C Borman. 1979. Personnel selection and classification systems. *Annual review of psychology* 30, 1 (1979), 477–525.

[34] Harrison Edwards and Amos J. Storkey. 2016. Censoring Representations with an Adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.*

[35] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 259–268.

[36] Society for Industrial, Organizational Psychology (US), and American Psychological Association. Division of Industrial-Organizational Psychology. 2018. *Principles for the validation and use of personnel selection procedures.* American Psychological Association.

[37] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning.* Springer series in statistics New York.

[38] Jim Fruchterman and Joan Melllea. 2018. *Expanding Employment Success for People with Disabilities.* Technical Report. benetech.

[39] Stacia Sherman Garr and Carole Jackson. 2019. *Diversity & Inclusion Technology: The Rise of a Transformative Market.* Technical Report. RedThread Research. https://info.mercer.com/rs/521-DEV-513/images/Mercer_DI_Report_Digital.pdf

[40] PW Gerhardt. 1916. Scientific selection of employees. *Electric Railway Journal* 47 (1916).

[41] 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 and Data Mining.* ACM.

[42] Jeff Grimmett. 2017. Veterinary Practitioners - personal characteristics and professional longevity. *VetScript* (2017).

[43] Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hannah Wallach, and Meredith Ringel Morris. 2019. Toward Fairness in AI for People with Disabilities: A Research Roadmap. *ACM SIGACCESS* 125 (October 2019).

[44] Richard A Guzzo, Alexis A Fink, Eden King, Scott Tonidandel, and Ronald S Landis. 2015. Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology* 8, 4 (2015), 491–508.

[45] Craig Haney. 1982. Employment tests and employment discrimination: A dissenting psychological opinion. *Indus. Rel. LJ* 5 (1982), 1.

[46] Kamala D. Harris, Patty Murray, and Elizabeth Warren. 2018. Letter to U.S. Equal Employment Opportunity Commission. https://www.scribd.com/embeds/388920670/content#from_embed

[47] Deborah Hellman. 2019. Measuring Algorithmic Fairness. *Virginia Public Law and Legal Theory Research Paper* 2019-39 (2019).

[48] Kimberly Houser. 2019. Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making. *Stanford Technology Law Review* 22 (2019).

[49] Amy E. Hurley-Hanson and Cristina M. Giannantonio (Eds.). 2016. Journal of Business Management. 22, 1 (2016).

[50] 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.* ACM, 49–58.

[51] Josh Jarrett and Sarah Croft. 2018. *The Science Behind The Koru Model of Predictive Hiring for Fit.* Technical Report. Koru.

[52] Stefanie K Johnson, David R Hekman, and Elsa T Chan. 2016. If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review* 26, 04 (2016).

[53] William F Kemble. 1916. Testing the fitness of your employees. *Industrial Management* (1916).

[54] Pauline T Kim. 2016. Data-driven discrimination at work. *Wm. & Mary L. Rev.* 58 (2016), 857.

[55] Pauline T Kim. 2017. Auditing algorithms for discrimination. *U. Pa. L. Rev. Online* 166 (2017), 189.

[56] Pauline T Kim. 2018. Big Data and Artificial Intelligence: New Challenges for Workplace Equality. *U. Louisville L. Rev.* 57 (2018), 313.

[57] Pauline T Kim. 2020. Manipulating Opportunity. *Virginia Law Review* 106 (2020).

[58] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics* 133, 1 (2017), 237–293.

[59] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2019. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* (2019).

[60] Robin SS Kramer and Robert Ward. 2010. Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology* 63, 11 (2010), 2273–2287.

[61] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.

[62] California State Legislature. 1959. Fair Employment and Housing Act.

[63] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Advances in Neural Information Processing Systems*. 8125–8135.

[64] George F Madaus and Marguerite Clarke. 2001. *The Adverse Impact of High Stakes Testing on Minority Students: Evidence from 100 Years of Test Data.* Technical Report. ERIC.

[65] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 3384–3393.

[66] Andrew Mariotti. 2017. *Talent Acquisition Benchmarking Report.* Technical Report. Society for Human Resource Management. https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/2017-Talent-Acquisition-Benchmarking.pdf

[67] Michael A Mcdaniel, Sven Kepes, and George C Banks. 2011. The Uniform Guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology* 4, 4 (2011), 494–514.

[68] Hugo Munsterberg. 1998. *Psychology and industrial efficiency.* Vol. 49. A&C Black.

[69] Isabel Briggs Myers. 1962. *The Myers-Briggs type indicator.* Consulting Psychologists Press.

[70] David Neumark, Roy J Bank, and Kyle D Van Nort. 1996. Sex discrimination in restaurant hiring: An audit study. *The Quarterly journal of economics* 111, 3 (1996), 915–941.

[71] Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology* 66, 6 (1963), 574.

[72] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM, 39–48.

[73] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 560–568.

[74] Ruchir Puri. 2018. Mitigating Bias in AI Models. *IBM Research Blog* (2018).

[75] 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.

[76] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *AAAI/ACM Conf. on AI Ethics and Society* (2019).

[77] McKenzie Raub. 2018. Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices. *Ark. L. Rev.* 71 (2018), 529.

[78] Lauren Rhue. 2018. Racial Influence on Automated Perceptions of Emotions. *Available at SSRN 3281765* (2018).

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

[80] John Roach. 2018. Microsoft improves facial recognition technology to perform well across all skin tones, genders. *The AI Blog* (2018).

[81] Michael C Rodriguez and Yukiko Maeda. 2006. Meta-analysis of coefficient alpha. *Psychological methods* 11, 3 (2006), 306.

[82] Edward Ruda and Lewis E Albright. 1968. Racial differences on selection instruments related to subsequent job performance. *Personnel Psychology* (1968).

[83] Eduardo Salas. 2011. Reply to Request for Public Comment on Plan for Retrospective Analysis of Significant Regulations pursuant to Executive Order 13563.

[84] Javier Sanchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to solve the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM.

[85] Heinz Schuler, James L Farr, and Mike Smith. 1993. *Personnel selection and assessment: Individual and organizational perspectives.* Psychology Press.

[86] Elaine W Shoben. 1978. Differential pass-fail rates in employment testing: Statistical proof under Title VII. *Harvard Law Review* (1978), 793–813.

[87] Elaine W Shoben. 1979. In defense of disparate impact analysis under Title VII: A reply to Dr. Cohn. *Ind. LJ* 55 (1979), 515.

[88] 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.

[89] Lewis Madison Terman. 1916. *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale.* Houghton Mifflin.

[90] Leona E Tyler. 1947. *The psychology of human differences.* D Appleton-Century Company.

[91] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management.* ACM, 22.

[92] John W Young. 2001. Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis. Research Report No. 2001-6. *College Entrance Examination Board* (2001).

[93] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, Vol. 54. PMLR, Fort Lauderdale, FL, USA, 962–970.

[94] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* ACM, 1569–1578.

[95] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning.* 325–333.

[96] Dawei Zhou, Jiebo Luo, Vincent MB Silenzio, Yun Zhou, Jile Hu, Glenn Currier, and Henry Kautz. 2015. Tackling mental health by integrating unobtrusive multimodal sensing. In *Twenty-Ninth AAAI Conference on Artificial Intelligence.*

A ADMINISTRATIVE INFORMATION ON VENDORS

Vendor name	Funding	# of employees	Location
8 and Above	–	1-10	WA, USA
ActiView	$6.5M	11-50	Israel
Assessment Innovation	$1.3M	1-10	NY, USA
Good&Co	$10.3M	51-100	CA, USA
Harver	$14M	51-100	NY, USA
HireVue	$93M	251-500	UT, USA
impress.ai	$1.4M	11-50	Singapore
Knockri	–	11-50	Canada
Koru	$15.6M	11-50	WA, USA
LaunchPad Recruits	£2M	11-50	UK
myInterview	$1.4M	1-10	Australia
Plum.io	$1.9M	11-50	Canada
PredictiveHire	A$4.3M	11-50	Australia
pymetrics	$56.6M	51-100	NY, USA
Scoutible	$6.5M	1-10	CA, USA
Teamscope	€800K	1-10	Estonia
ThriveMap	£781K	1-10	UK
Yobs	$1M	11-50	CA, USA

Table 3: Administrative information