Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices

Manish Raghavan* Solon Barocas† Jon Kleinberg* Karen Levy*

Abstract

There has been rapidly growing interest in the use of algorithms for employment assessment, especially as a means to address or mitigate bias in hiring. Yet, to date, little is known about how these methods are being used in practice. How are algorithmic assessments built, validated, and examined for bias? In this work, we document and assess the claims and practices of companies offering algorithms for employment assessment, using a methodology that can be applied to evaluate similar applications and issues of bias in other domains. 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 techniques for detecting and mitigating bias. We find that companies' formulation of "bias" varies, as do their approaches to dealing with it. We also discuss the various choices vendors make regarding data collection and prediction targets, in light of the risks and trade-offs that these choices pose. We consider the implications of these choices and we raise a number of technical and legal considerations.

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. At first glance, it might appear that social and academic pressure for companies to consider normative goals when building algorithms has led to positive industry change in these areas, and indeed, many companies have publicly responded to these calls for improvement.[1] 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 and metrics of fairness allow us to ask focused questions about the meaning of "fair" and "unbiased" models. Thus, rather than viewing these as a way to "solve" bias, we see these definitions and metrics as providing a concrete way to analyze the claims and practices of industry: computational approaches to fairness give us a framework to guide empirical research about companies' practices.

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, 12]. Indeed, when these models are

*Cornell University

†Microsoft Research and Cornell University

[1] See, e.g., I.B.M.'s "Diversity in Faces" project [73] and Microsoft's response to critical research [61].

available, we can and should investigate them to identify potential points of concern. But what do we do when we have little or no access to models or the data that 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 place users' privacy at risk. As we study 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 inaccessible to researchers and the general public. 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 if we do not have 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. A great deal of work on algorithmic fairness considers hiring as a motivating example of the risks of discrimination that warrant careful attention. Despite this, we have little concrete information about how industry has handled these concerns when adopting algorithmic techniques. In this work, we seek to shed further light on algorithmic hiring by examining vendors of algorithmic pre-employment assessments, who provide tools to quantitatively evaluate job-seekers for hiring purposes. Following a review of firms offering recruitment technologies, we identify 19 vendors of pre-employment assessments. We document what each company has disclosed about their practices and consider the implications of these claims. In so doing, we develop an understanding of how attempts to address bias have take place in industry practice and what critical issues these have been left unaddressed.

Prior work has sought to taxonomize the points at which bias can enter machine learning systems [5, 47]. Barocas and Selbst describe how the choice of target variable, collection of training data, labeling of examples, and measurement of features are all potential sources of disparities [5]. Similarly, Kleinberg et al. note that discrimination can result from the choice of outcome, choice of features, or choice of training algorithm [47]. 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 regards 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 an 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 from the information we consider.

We stress that our analysis is not intended as an exposé of industry practices. Many of the vendors we study exist precisely because their founders seek to provide a fairer alternative to traditional hiring practices known to be problematic. Our hope is that this work will paint a realistic picture of the landscape of algorithmic techniques in pre-employment assessment and offer some 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 the legal precedents surrounding them. In Section 3, we systematically review vendors of algorithmic screening tools and provide empirical findings on their practices based on the claims that they make. We analyze these practices in detail in Section 4, examining particular causes for concern and providing recommendations. We provide concluding thoughts in Section 5.

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. The Society for Human Resource Management estimates that the average time to fill a position in 2016 was over a month [52]. As a result, many employers seek tools with which to optimize their hiring processes.

The hiring pipeline consists of a series of stages leading to offers being made to chosen candidates. Broadly speaking, there are four distinct stages, though the boundaries between them are not always rigid: sourcing, screening, interviewing, and selection [10]. 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. Our work will focus on the *screening* stage, and in particular, pre-employment assessments that use algorithmic techniques to assess 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, along with the concerns that they raise for fairness. For example, Bogen and Rieke provide an overview of the various ways in which algorithms are being introduced into this pipeline, with a particular focus on their implications for equity [10]. Garr surveys a number of platforms designed to promote diversity and inclusion in hiring [34]. Broadly considering the use of data science in HR-related activities, Cappelli et al. identify a number of challenges and propose a framework to help address them [14]. Ajunwa provides a legal framework to consider the problems algorithmic tools introduce and argues against subjective targets like "cultural fit" [1].

It is important to note that even in the absence of algorithms, the hiring process is widely acknowledged to be fraught with bias. A well-known study by Bertrand and Mullainathan demonstrated that given identical resumes, employer response rates were significantly higher when candidates had names that suggested they were white males as compared to other groups [9]. This results of this study, and others like them [8, 7, 43], are widely accepted in the world of human resources, where practitioners continually seek new ways of handling bias [49]. Advocates argue for algorithmic techniques as a means to address bias [16, 27], and indeed, there is some evidence that they can be used to combat human idiosyncrasies [40, 46]. However, while these tools have the potential to mitigate certain human biases, they run the risk of reinforcing or creating new inequalities as well.

A history of equity concerns in assessment. Pre-employment assessments have a long history, beginning with examinations for the Chinese civil service thousands of years ago [38]. In the early 1900's, the idea that assessments could reveal innate cognitive abilities gained traction in both industrial and academic circles, leading to the formation of Industrial Psychology as an academic discipline [54, 35, 44]. During the two World Wars, the US government turned to these assessments in an attempt to quantify the abilities of its soldiers, paving the way for their widespread adoption in postwar industry [4, 30, 31]. Historically, these assessments were primarily behavioral or cognitive in nature, like the Stanford-Binet IQ test [74], the Myers-Briggs type indicator [55], and the Big Five personality traits [57]. Industrial-Organizational (IO) Psychology remains a prominent component of these modern assessment tools—many vendors we examine employ a team of IO psychologists who work in concert with data scientists to create and validate assessments.

More recently, scholars in the field of 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 [37]. 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 [15]. 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 [26].

This is particularly worrying because cognitive assessments have imposed adverse impacts on minority populations since their introduction into mainstream American use [75, 67, 25]. Critics have long contended that observed group differences in test outcomes indicated flaws in the tests themselves [28], 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 [25].[2]

In light of these concerns, the American Psychological Association (APA) includes appeals to fairness and bias in its Principles for the Validation and Use of Personnel Selection Procedures [32]. Recognizing that "fairness" is ill-defined and means different things to different stakeholders, the APA instead cautions against predictive bias: systematic errors in predictions for a certain group [32]. Moreover, while the APA Principles strive to equalize *opportunity* for candidates of all backgrounds, they explicitly reject equalizing *outcomes* in the form of "equal passing rates for subgroups of interest" [32]. As we will see, this rejection of outcome-based notions of bias 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") and created the Equal Employment Opportunity Commission (EEOC) to ensure compliance [21]. 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 [20], 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 the predictive ability of an assessment, and demonstrating criterion validity entails demonstrating that 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. Finally, assessments demonstrate construct validity if they test for some construct (e.g., grit or leadership) that is required for good job performance. This, so far, is in keeping with APA Principles, which place a similar emphasis on validity [32].

When is an assessment legally considered discriminatory? Based on existing precedent, the

[2] Disparities in assessment outcomes for minority populations are not limited to pre-employment assessments. In the education literature, the adverse impact of assessments on minorities is well-documented [51]. This has led to a decades-long line of literature seeking to measure and mitigate the observed disparities (see [41] for a survey).

Uniform Guidelines provide two avenues to challenge an assessment: disparate treatment and disparate impact [5]. Disparate treatment is relatively straightforward—it is illegal to explicitly treat candidates differently based on categories protected under Title VII [20, 21]. Disparate impact is more nuanced, and while we provide an overview of the process here, we refer the reader to [5] for a more complete discussion.

Under the Uniform Guidelines, a case can be brought against an employer for disparate impact if the selection rate for one protected group is less than $^4/_5$ of that of another group [20]. Once this disparity is established, an employer must respond by showing that the selection procedures that it uses are both valid and necessary from a business perspective [20]. Moreover, an employer can be held liable if the plaintiff can show the existence of an alternative selection procedure with less adverse impact that the employer could have used instead with little business cost [20].[3] Importantly, the Title VII and EEOC requirement of approximately equal selection rates via the $^4/_5$ rule conflicts with the APA Principles, although the APA Principles do point out that "group differences [in selection rates] should trigger heightened scrutiny for possible sources of bias" [32]. Responding to this inconsistency, psychologists have argued that the Uniform Guidelines "substantially deviate from scientific knowledge and professional practice" and should be revised in light of more recent research [53].

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. We identified 322 companies providing these assessments by combining the top 300 start-up companies (by funding amount) on Crunchbase under its "recruiting" category[4] with an inventory of relevant companies found in reports by Upturn [10] and RedThread Research [34]. 39 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. 21 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 19 vendors in our sample. With these 19 vendors, in April 2019, 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 [5, 47], we ask the following questions of vendors:

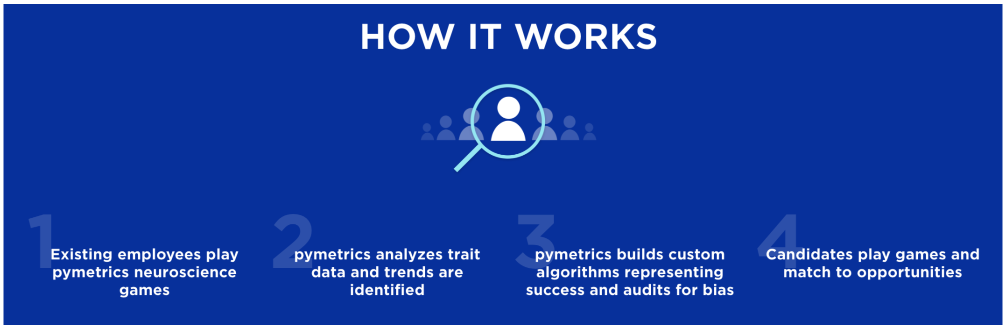
[3] It should be noted that this description is based on a particular (although the most common) interpretation of Title VII. Legal scholars contend that Title VII may offer stronger protections to minorities [11, 45], and there is disagreement on how (or whether) to operationalize the $^4/_5$ rule through statistical tests [70, 18, 71, 19]. For the purposes of this work, we will not consider alternative interpretations of Title VII, nor will we get into the specifics of how exactly violations of the $^4/_5$ rule should be detected.

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

- What types of assessments do they provide? (Questions? Videos? What are the features?)
- What is the target variable they aim to predict (e.g., sales revenue or employee evaluations)?
- Where do the training data come from? Do they train on data from the client, or do they use their own data sources?
- What information do they provide on the validation of their assessments? Do they release validation studies or whitepapers?
- What concrete claims or guarantees (if any) do they provide regarding discrimination? When applicable, how do they achieve these guarantees?

To answer these questions, we exhaustively searched the website of each company. This included 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 companies.

3.2 Findings



HOW IT WORKS

1 Existing employees play pymetrics neuroscience games

2 pymetrics analyzes trait data and trends are identified

3 pymetrics builds custom algorithms representing success and audits for bias

4 Candidates play games and match to opportunities

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

In our review, we found 19 vendors providing algorithmically driven pre-employment assessments. Those that had available funding information on Crunchbase (17 out of 19) ranged in funding from around $1 million to $93 million. Most vendors (15) had 50 or fewer employees, and roughly half (9) were based in the United States. 16 vendors were present in Crunchbase's "Recruiting Startups" list; the remaining vendors were taken from reports by Upturn [10] and RedThread Research [34]. 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 (12 vendors), video interview analysis (6 vendors), and gameplay (e.g., puzzles or video games) (6 vendors). Note that many vendors (e.g., HireVue) 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.

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

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. In the "Custom?" column, C denotes "custom" (uses employer data), S denotes "semi-custom" (qualitatively tailored to employer without data) and P denotes "pre-built." In the "Adverse impact" column, the recorded phrases were found on the websites of the vendors in question.

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.

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. 8 of these 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 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 predetermined 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 produce scores based on a small number of questions, modern assessments differ in that they may build a psychographic profile by using machine learning to analyze a rich data source like a video or gameplay.

Validation. Very few vendors provide concrete information on the validation of their assessments, although some vendors claim to validate their models without providing details. Good & Co.,[5] notably, provides fairly rigorous validation studies of the psychometric component to 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 14 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. 3 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 their models for bias, “fixing” it when it appeared. HireVue, in particular, offered a detailed description of their approach to de-biasing, which involves removing features correlated with protected attributes when adverse

[5] https://good.co/

impact is detected. Other vendors (e.g., pymetrics, Knockri, and PredictiveHire) claimed to "fix" adverse impact when it is found without going into the details of how they do this.

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 some underlying cognitive or behavioral traits, so the outcome of an assessment is a small number of scores, one for each competency being measured. These scores can then be combined to form a single number (often known as a "fit" score) based on the competencies deemed necessary for the particular role being selected for. 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" [42]. When a vendor claims that an assessment like this is naturally unbiased, it means that the distribution of scores is similar across demographic groups.

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: build a model and test it for adverse impact against various subgroups.[6] 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 downweights or removes features found to be highly correlated with the protected attribute in question, noting that this can significantly reduce adverse impact while having little effect on the predictive accuracy of the assessment. In Section 4.3, we will provide an in-depth discussion of these efforts to define and guarantee the removal of bias.

4 Analysis of Technical Concerns

Our findings in Section 3 raise several concerns about the pre-employment assessment process. While not an exhaustive list of the myriad vectors for algorithmic bias, we focus our attention on three particular areas identified in our empirical review. Creating assessments via ML requires the collection of data from which a predictive model can be learned. In order to do so, a vendor must make several **data choices**—for example, they may train models to predict success based on the employees the client has chosen to hire in the past. Vendors may choose to use **alternative assessment formats** like game-based assessments or analysis of recorded video interviews that produce far more features and require more complex ML tools than traditional question-based assessments. Finally, many vendors take steps to **detect or remove bias** in their assessments, leading to various "de-biasing" methodologies. In the remainder of this section, we analyze vendors with regard to these three practices.

Our analysis does not intend to suggest that all or even most of the vendors surveyed are handling these concerns poorly; many actively think about and try to address them. However, we find it useful to survey some of the issues that may arise when machine learning techniques are used to attempt to predict job performance, especially since it is often not obvious *which* vendors are particularly diligent when it comes to preventing biased 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 in a learning pipeline. Where do the

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

data come from? What is the "right" outcome to predict? These and others are crucial decisions in the ML pipeline, and their impacts on the bias of a system should not be discounted.

Custom assessments. Consider a hypothetical practitioner who sets out to create a custom assessment to determine who the "best" candidates are 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 [59]. It quickly becomes clear that she must somehow 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 [14]. 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 no way to get 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 assessment performance to their observed job performance. How many employees does she need data from in order to get an accurate model? What if certain employees don't want to or don't have time to take the assessment? Might there be some sort of response bias? Is the set of employees who respond representative of the larger applicant pool who will ultimately be judged based on this assessment?

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, have been found to display biases towards minorities [72, 56, 64]. Moreover, predicting the success of future employees based on current employees inherently skews the task towards finding candidates who resemble 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 tools for employers to find candidates who are compatible with members of a current team. Good & Co. further advertises their assessments as a way to "[r]eplicate your top performers."[7] When models are fit and 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. In principle, any role at any company could have its own version of a predictive model, tailor-made for the particular team a candidate would be joining. Does each one need to be audited for bias? How would a vendor go about doing so?

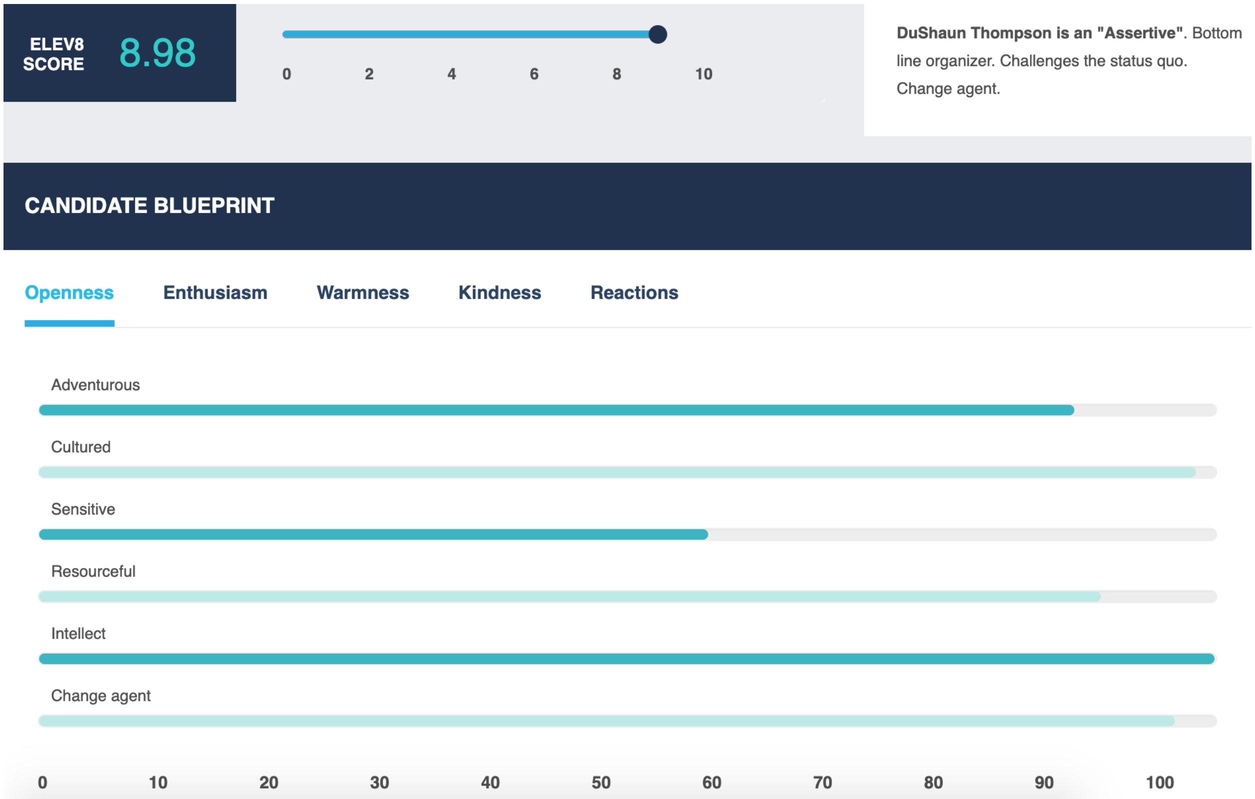
And yet, while it is easy to criticize vendors for the choices they make, it's not clear that there are better alternatives. In practice, it is impossible to even define, let alone collect data on, an objective measure of a "good" employee. Nor is it always feasible to get data on a completely representative sample of candidates. Vendors and advocates argue 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 [16].

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 type of job) that has been validated across data from a variety of clients. This has the

[7] `https://good.co/pro/`

advantage that it isn't subject to the idiosyncratic data of each client, and moreover, it can draw from a more diverse range of candidates and employees to learn a broader notion of what a "good" employee looks like. Additionally, pre-built assessments may be attractive to clients who do not have enough existing employees from whom a custom assessment can be built.

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 reports a single "Elev8" score tailored to the particular client.



ELEV8 SCORE 8.98

DuShaun Thompson is an "Assertive". Bottom line organizer. Challenges the status quo. Change agent.

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`)

Necessary trade-offs. 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 [66], it may be difficult to create an objective benchmark 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 [32], and both underlying candidate pools and job specifics differ across locations, companies, and job descriptions. Pre-built assessments are not trained directly on a client's outcome data, and therefore may not adapt well to the particular requirements a client has.

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 [6, 33]. 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 creators of assessments in this process [32]; 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 [15]. Game- and video-based assessments, in particular, are becoming 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 [50, 36] as evidence that these assessments truly do 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 their predictive validity, it is important to bear in mind that 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 truly important 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 is indicative of 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 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" [32]. Yet when an algorithm takes in "millions of data points" for each candidate (as advertised by pymetrics[8]), 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 the use of a problematic relationship, especially because rich sources of data can easily encode or correlate with properties that are unethical or illegal to use in the hiring process. Facial analysis,

[8] `https://perma.cc/3284-WTS8`

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 [12, 62, 63]. 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[9]) 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 [39]. 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 [61, 65], 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 [50, 77]. Moreover, for ethical and legal reasons, vendors for the most part cannot and should not collect information about attributes like candidates' health [22], making it difficult or even impossible to detect whether the relatively opaque deep learning models used for facial analysis inadvertently learn proxies for prohibited features.

4.3 Detecting and Removing Bias

Even when an assessment has been validated, it still has the potential to lead to biased outcomes. To this end, many vendors take steps to ensure that their assessments don't display certain forms of bias. Here, we evaluate how vendors detect and mitigate bias, placing their techniques in the context of bias-removal efforts in related domains.

Inherent challenges in defining unbiased assessments. Formally defining what it means for an assessment to be biased is an intrinsically difficult task. It is impossible to separate notions of fairness and bias from the politics and values that underpin them. Different stakeholders have different conceptions of what it means to be fair, and these are not always compatible with one another [17, 48].

For example, the EEOC Guidelines' push towards minimizing avoidable inequalities in outcomes is at odds with the APA Principles, which define unbiased assessments as those that have neither differential validity (disparities in accuracy between subgroups) nor differential prediction (where the optimal mapping from features to predictions differs between subgroups) [20, 32]. Given these competing conceptions of fair assessment, it is not obvious how the tension between them should be resolved. Beyond this disagreement in goals, classical techniques used to examine bias in assessments typically consider the case where scores are determined using regression [32, 76], which is much more tractable to analyze than more modern machine learning techniques.

Moreover, normative decisions regarding the formalization of bias are subject to practical and technical constraints. 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 [20, 32]. The same holds true for claims regarding bias: disparities in outcomes (or the lack thereof) may depend on whether the assessment is being taken

[9] `https://www.affectiva.com/`

by recent college grads in Michigan applying for sales positions or high school dropouts in New York applying for jobs stocking warehouses. It may be especially hard to find a representative sample when minority populations are small—does a vendor have enough sample videos of individuals with disabilities to verify that its algorithms aren't (perhaps inadvertently) discriminatory? And even if they do, is it possible for applicants or government regulators to challenge such claims? Again, there is no obvious answer here: certain populations are inherently small or hard to identify, making it challenging both for vendors seeking to provide quality service and for regulators seeking to protect marginalized groups. Thus, while it is tempting to look for simple technical definitions of fairness and bias, there are significant barriers to implementing even simple constraints in practice.

With these challenges in mind, consider the role of our hypothetical data scientist in producing unbiased assessments. How does she operationalize abstract notions of fairness and equity? What values is she implicitly building into the system? And, importantly, to what extent are her choices shaped by financial and legal constraints?

As noted in Section 3, several vendors guarantee that their assessments are "unbiased" by building a model, testing it for adverse impact (using statistical tests aimed at the 4/5 rule), and modifying the weights or features used until they are satisfied. From a vendor's perspective, this strict adherence to the 4/5 rule can be seen as a logical business decision: given the legal uncertainty surrounding machine learning in the context of hiring, both vendors and their clients prefer to steer clear of any regulatory trouble. In order to better understand how and why this variant of algorithmic de-biasing has come to be, it is useful to place it in the context of how bias is handled in other domains.

De-biasing in related domains. Questions regarding the measurement and removal of bias are by no means unique to algorithmic hiring; to our knowledge, however, the particular approaches taken by vendors of algorithmic pre-employment assessments are not widely used in other contexts. To shed some light as to why this is the case, we examine practices from lending, educational testing, and historical methods from pre-employment assessments for comparison. In this context, we will consider the justification for algorithmic de-biasing techniques used by vendors like HireVue and pymetrics and discuss some of the legal questions they raise.

In lending, creditors use statistical models to determine who to offer loans. Discrimination in lending with respect to protected attributes (e.g., race, color, religion, national origin, and sex) is prohibited by the Equal Credit Opportunity Act (ECOA) [13]. To address any potential disparate impact generated by their use of statistical models, creditors attempt to justify their selection procedures by inspecting each feature in their models and developing some qualitative defense for its inclusion. In particular, lenders often seek to ensure that there is an "understandable relationship" between a given feature and creditworthiness [58] — a story that lenders can tell to explain a feature's relevance. This approach permits objections to the use of certain features, regardless of their predictive value; at the same time, it allows for the use of qualitatively justified features that may result in avoidably unequal selection rates.

In educational testing, Scheuneman introduced the idea of Differential Item Functioning (DIF), which defined an item (i.e., test question or section) as unbiased if, conditioned on the outcome target of the tests, different groups had equal performance on the item [68]. Thus, an item demonstrates DIF if individuals from different subgroups with similar overall scores perform differently on that particular item [29]. Under this definition, individual items could be analyzed for bias and modified or removed [60]. Similarly, algorithmic assessment vendors consider individual features to control biased outcomes; however, they differ in that vendors target features correlated with a protected attribute, while DIF looks for such correlation *conditioned on "ability,"* as measured by the test itself. Crucially, this distinction implies that DIF still allows for group differences in outcome: score distributions or selection rates can look arbitrarily different for a test that is "unbiased" under DIF, as long as those differences can be explained by disparities in underlying qualification or ability.[10] Assessment vendors engaged in the active de-biasing we observe, on the other hand, explicitly control the outcomes they produce, effectively scrubbing their data of correlations to protected attributes until outcomes are equalized to within a tolerable range.

Finally, we find it instructive to consider the methods used in the General Aptitude Test Battery (GATB), a pre-employment assessment developed in the 1940s by the US Employment Service to match job-seekers to employers. The GATB was quickly found to have an adverse impact on ethnic minorities [69]. In response, results were reported as within-group percentile scores by ethnicity—black, Hispanic, and other—instead of raw scores [25]. A National Academy of Sciences study was commissioned to consider, among other factors, the justification for such a policy. The report found evidence of both differential validity and differential prediction; that is, both the predictive validity of the test and the optimal scoring rule (to produce a score from the raw responses) differed across racial groups [25]. Moreover, the report noted that for a given level of job performance, minority candidates tended to perform worse on the test [25]. As a result, they found that without within-group reporting, minority applicants would suffer from "higher false-rejection rates" [25], leading them to recommend the continued reporting of within-group percentiles. Note that unlike the techniques used in lending and educational testing, the GATB's within-groups percentile reporting was designed to equalize outcomes: by definition, equal proportions of each subgroup would be above any particular percentile score.

The design of algorithmic de-biasing. In principle, any of the above techniques (and surely many more from other domains) could have been used to mitigate bias in algorithmic pre-employment assessment. Why did vendors settle upon adverse impact testing and post-hoc corrections? Part of the answer lies in the fact that many modern assessments include tens of thousands of features per applicant, and it would be infeasible to manually inspect and justify each one as done in lending. Even more salient is the fact that while the 4/5 rule is by no means a hard constraint, vendors have an incentive to treat it as such: their clients do not want to risk running afoul of Title VII and EEOC guidelines. Thus, although inspecting individual inputs for bias, as is done in lending or educational testing, could be legally compliant, vendors face pressure to control the selection rates of the assessments they produce to satisfy the 4/5 rule.

In principle, within-group reporting of scorse as found in the GATB would lead to compliance with the 4/5 rule, and as some computer scientists have argued, doing so may be the optimal way to equalize selection rates [24]; so why don't vendors use it? In fact, there are strong technical, legal, and political reasons not to do so. From a technical perspective, it is significant that protected attributes are *intersectional*: an individual can belong to multiple protected groups based on race, gender, age, and other attributes. Because racial discrepancies observed in the GATB were so salient, administrators chose to report within-group percentiles by race, but if outcomes were to vary significantly by gender or age as well, vendors might need to report percentiles within a large number of intersectional subgroups. Perhaps more importantly, within-group reporting would likely be considered illegal today. In 1986 the Department of Justice challenged the legality of within-group scoring in employment assessment on the grounds that it constituted explicitly disparate treatment [69]. While the U.S. National Research Council found that within-group scoring was necessary to prevent unjustifiable adverse impact in the GATB, the practice was effectively outlawed

[10] In the extreme, a test can pass all candidates from one group and none from another and still be unbiased as measured by DIF.

by the Civil Rights Act of 1991 [23]. Moreover, vendors may also find it politically infeasible to adopt such a solution, as it would effectively constitute an admission that the underlying assessment performs quite differently for different subgroups.

Implications. Despite the perhaps good reasons not to follow precedents for the detection and removal of bias set in a number of related domains, there are still important consequences of algorithmic de-biasing that are worth considering.

First, 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 [32, 76]. But because of the legal emphasis placed on adverse impact, vendors have little incentive to structure their techniques around it.

Moreover, bias is not limited to the task of predicting outputs from inputs. Vendors should critically examine the system of producing an assessment as a whole. 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, we believe that the public emphasis on the 4/5 rule as a definition of bias runs the risk of downplaying the importance of examining a system as a whole, as opposed to ensuring the narrow property that selection rates should be roughly equal.

From a policy perspective, the EEOC can and should clarify its position on the use of algorithmic de-biasing techniques. As of now, their legality is unclear, and they have yet to be challenged to our knowledge. While existing guidelines can be argued to apply to ML-based assessments, the de-biasing techniques described above do present new opportunities and challenges. Suppose, for instance, that a vendor supplies a model that results in adverse impact. It might argue that the model is properly validated, and that its use constitutes a business necessity. If, however, the vendor could have reduced the adverse impact through algorithmic de-biasing without significantly reducing predictive ability, should this be considered an "alternative business practice" and therefore render the vendor or employer liable for not using it? And if so, does this imply that algorithmic de-biasing should be required for all vendors? Existing guidelines do not answer these questions, and both vendors and candidates could benefit from the EEOC taking a concrete stance on de-biasing techniques.

5 Discussion

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 infeasible to perform a traditional audit; despite this, we are able to glean valuable information simply 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 models and practices. This analysis demonstrates the value of **transparency**: the more vendors disclose about their practices, the more confidently we can assess the models that they are likely to produce. From a vendor's perspective, they have the ability to shape industry standards and best practices through the information that they disclose publicly, and they have an incentive to ensure that the industry as a whole remains both legally and socially acceptable.

In analyzing models via practices, we observe that it is both useful and important 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.

Finally, 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 [10, 34]) 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 uses of technology in the hiring pipeline (see, e.g., Textio[11] or Jopwell[12]), it allowed us to make specific and direct comparisons between vendors and get a more detailed understanding of the technical challenges specific to assessments.

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?). We believe that none of these questions can be completely addressed without drawing from a broad range of perspectives. Because the systems we examine are shaped by technical, legal, political, and social forces, taking an interdisciplinary view allows us 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, Miranda Bogen, Heather Bussing, Albert Chang, A. F. Cooper, Fernando Delgado, Kate Donahue, Stacia Garr, Jim Guszcza, Stephen Hilgartner, Lauren Kilgour, Loren Larsen, Cassidy McGovern, Helen Nissenbaum, Samir Passi, Sarah Riley, David Robinson, Caleb Rottman, John Sumser, Katie Van Koevering, Briana Vecchione, Angela Zhou, Malte Ziewitz, and Lindsey Zuloaga for their suggestions and insights.

References

[1] Ifeoma Ajunwa. The paradox of automation as anti-bias intervention. *Cardozo Law Review*, Forthcoming.

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

[3] Illinois General Assembly. Video interview act, 2019.

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

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

[11] Textio (`https://textio.com/`) analyzes job descriptions for gender bias and makes suggestions for alternative, gender-neutral framings.

[12] Jopwell (`https://www.jopwell.com/`) builds and maintains a network of Black, Latinx, and Native American students and partners with employers to connect students to opportunities.

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

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

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

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

[10] Miranda Bogen and Aaron Rieke. Help wanted - an exploration of hiring algorithms, equity, and bias. Technical report, Upturn, 2018.

[11] Stephanie Bornstein. Antidiscriminatory algorithms. *Ala. L. Rev.*, 70:519, 2018.

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

[13] Bureau of Consumer Financial Protection. Equal credit opportunity act, 2011.

[14] Peter Cappelli, Prasanna Tambe, and Valery Yakubovich. Artificial intelligence in human resources management: Challenges and a path forward. *Available at SSRN 3263878*, 2018.

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

[16] Tomas Chamorro-Prezumic and Reece Akhtar. Should companies use ai to assess job candidates? *Harvard Business Review*, 2019.

[17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[18] Richard M Cohn. On the use of statistics in employment discrimination cases. *Ind. LJ*, 55:493, 1979.

[19] Richard M Cohn. Statistical laws and the use of statistics in law: A rejoinder to Professor Shoben. *Ind. LJ*, 55:537, 1979.

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

[21] U.S. Congress. Civil rights act, 1964.

[22] U.S. Congress. Americans with disabilities act, 1990.

[23] U.S. Congress. Civil rights act, 1991.

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

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

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

[27] Bo Cowgill. Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University*, 29, 2018.

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

[29] Neil J Dorans and Paul W Holland. DIF detection and description: Mantel-Haenszel and standardization 1, 2. *ETS Research Report Series*, 1992(1):i–40, 1992.

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

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

[32] Society for Industrial, Organizational Psychology (US), and American Psychological Association. Division of Industrial-Organizational Psychology. *Principles for the validation and use of personnel selection procedures.* The Society, 2003.

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

[34] Stacia Sherman Garr and Carole Jackson. Diversity & inclusion technology: The rise of a transformative market. Technical report, RedThread Research, 2019.

[35] PW Gerhardt. Scientific selection of employees. *Electric Railway Journal*, 47, 1916.

[36] Jeff Grimmett. Veterinary practitioners - personal characteristics and professional longevity. *VetScript*, 2017.

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

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

[39] Kamala D. Harris, Patty Murray, and Elizabeth Warren. Letter to U.S. Equal Employment Opportunity Commission, 2018.

[40] Kimberly Houser. Can ai solve the diversity problem in the tech industry? mitigating noise and bias in employment decision-making. *Mitigating noise and bias in employment decision-making (February 28, 2019)*, 22, 2019.

[41] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. *arXiv preprint arXiv:1811.10104*, 2018.

[42] Josh Jarrett and Sarah Croft. The science behind the Koru model of predictive hiring for fit. Technical report, Koru, 2018.

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

[44] William F Kemble. Testing the fitness of your employees. *Industrial Management*, 1916.

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

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

[47] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 2019.

[48] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science*, 2017.

[49] Rebecca Knight. 7 practical ways to reduce bias in your hiring process. *Harward Business Review*, 2017.

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

[51] George F Madaus and Marguerite Clarke. The adverse impact of high stakes testing on minority students: Evidence from 100 years of test data. Technical report, ERIC, 2001.

[52] Andrew Mariotti. Talent acquisition benchmarking report. Technical report, Society for Human Resource Management, 2017.

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

[54] Hugo Munsterberg. *Psychology and industrial efficiency*, volume 49. A&C Black, 1998.

[55] Isabel Briggs Myers. *The myers-briggs type indicator*. Consulting Psychologists Press, 1962.

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

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

[58] Office of the Comptroller of the Currency. Fair lending. Technical report, Comptroller of the Currency Administrator of National Banks, 2010.

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

[60] Randall D Penfield. Fairness in test scoring. In *Fairness in Educational Assessment and Measurement*, pages 71–92. Routledge, 2016.

[61] Ruchir Puri. Mitigating bias in AI models. *IBM Research Blog*, 2018.

[62] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *AAAI/ACM Conf. on AI Ethics and Society*, 2019.

[63] Lauren Rhue. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765*, 2018.

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

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

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

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

[68] Janice Scheuneman. A method of assessing bias in test items. *Journal of Educational Measurement*, 16(3):143–152, 1979.

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

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

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

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

[73] John R. Smith. Ibm research releases 'diversity in faces' dataset to advance study of fairness in facial recognition systems. *IBM Research Blog*, 2019.

[74] Lewis Madison Terman. *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, 1916.

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

[76] John W Young. Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis. Research report no. 2001-6. *College Entrance Examination Board*, 2001.

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

A Administrative Information on Vendors

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

Table 3: Administrative information