Don't let Ricci v. DeStefano Hold You Back: A Bias-Aware Legal Solution to the Hiring Paradox

Jad Salem*
Deven R. Desai*
Swati Gupta*
jsalem7@gatech.edu
deven.desai@scheller.gatech.edu
swatig@gatech.edu
Georgia Institute of Technology
Atlanta, USA

ABSTRACT

Companies that try to address inequality in employment face a hiring paradox. Failing to address workforce imbalance can result in legal sanctions and scrutiny, but proactive measures to address these issues might result in the same legal conflict. Recent run-ins of Microsoft and Wells Fargo with the Labor Department's Office of Federal Contract Compliance Programs (OFCCP) are not isolated and are likely to persist. To add to the confusion, existing scholarship on Ricci v. DeStefano often deems solutions to this paradox impossible. Circumventive practices such as the 4/5ths rule further illustrate tensions between too little action and too much action.

In this work, we give a powerful way to solve this hiring paradox that tracks both legal and algorithmic challenges. We unpack the nuances of Ricci v. DeStefano and extend the legal literature arguing that certain algorithmic approaches to employment are allowed by introducing the legal practice of banding to evaluate candidates. We thus show that a bias-aware technique can be used to diagnose and mitigate "built-in" headwinds in the employment pipeline. We use the machinery of partially ordered sets to handle the presence of uncertainty in evaluations data. This approach allows us to move away from treating "people as numbers" to treating people as individuals—a property that is sought after by Title VII in the context of employment.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Mathematics of computing** → **Discrete mathematics**.

KEYWORDS

anti-discrimination laws, hiring, resume screening, bias, uncertainty

*All authors contributed equally to this work.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533129

ACM Reference Format:
Jad Salem, Deven R. Desai, and Swati Gupta. 2022. Don't let Ricci v. DeStefano Hold You Back: A Bias-Aware Legal Solution to the Hiring Paradox. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3531146.3533129

1 INTRODUCTION

How employers identify whom to interview and then hire has important effects across society. Employment significantly affects access to healthcare, continuing education and, therefore, quality of life. The benefits of employment are not, however, evenly distributed across race and gender categories in the United States. After George Floyd's death, companies acted to address racial injustice by making public statements, donations to support racial equality, and Juneteenth a company holiday [52]. Several companies went further. Microsoft announced a $150 million investment to improve diversity including setting a goal of doubling the number of "Black and African American people managers, senior individual contributors and senior leaders" in the United States by 2025 [53]. Wells Fargo made a commitment to "double Black Leadership" by 2025 and "will evaluate senior leaders based on their progress in improving diversity and inclusion in their areas of responsibility, in addition to other efforts" [53]. Google has set a goal of having 30% of its leadership from "under represented groups" by 2025 [34]. Yet, both Microsoft and Wells Fargo received letters from the Labor Department's Office of Federal Contract Compliance Programs (OFCCP) due to concern that the plans may discriminate based on race [52]. At the same time, the OFCCP announced a settlement with Microsoft in September 2020 for $3 million back pay and interest to address hiring disparities "against Asian applicants" for several positions from December 2015 to November 2018 [85]. The two OFCCP positions clash and appear to create a world where inaction opens the company to litigation, if not breaking the law, and corrective action creates the same risks. One might argue that the recent OFCCP inquiries were peculiar to the Trump administration's approach to this area of law and not something the current administration would pursue. Administrations, however, change and a new one might follow the Trump approach. Regardless of who is in the White House, legal activism to challenge steps taken to

address diversity or challenge discriminatory results are not likely to go away.[1]

A company may pursue diversity goals and/or address affirmative action plans, but the two are not the same, and the difference matters [57]. As the EEOC explains in "Section 15 Race and Color Discrimination" of its Compliance Manual, diversity can be understood as "a business management concept under which employers voluntarily promote an inclusive workplace" [18]. Companies have pursued diversity to attract talent and gain "a competitive advantage" [18]. In contrast, affirmative action refers to "those actions appropriate to overcome the effects of past or present practices, policies, or other barriers to equal employment opportunity" [4]. Such steps may occur because of a court order, negotiated settlement, or government regulation [18]. Employers may also use a voluntary affirmative action plan "in appropriate circumstances, such as to eliminate a manifest imbalance in a traditionally segregated job category" [18]. There is a conceptual and practical link between diversity goals and affirmative action. A company may pursue diversity "for competitive reasons rather than in response to discrimination" and "such initiatives may also help to avoid discrimination" [18]. As the legal status of diversity plans is unclear, methods to support both options are needed.

As another motivation, companies may want to see whether they are missing hiring and talent opportunities. Companies can be stuck in an equilibrium because they rely on, or exploit "old certainties," rather than explore "new possibilities" [81]. This exploration/exploitation trade-off began in organizational business literature but has become a significant part of how the machine learning community thinks about understanding information [51]. As a matter of best organizational and ML practices, companies need ways to explore new candidate pools.

Regardless of the motivation behind a company plan, there is a steady drumbeat for algorithmic transparency, especially in employment and admissions contexts [47]. Thus an entity may have to or wish to reveal their process to show that it is mathematically and legally sound. These issues could push a company to avoid steps to address diversity because of real or perceived litigation risks. Although some argue that the use of machine learning would constitute a valid business necessity claim so long as the target variable is job-related (thus rendering the question of equality of outcomes irrelevant) debates about which actions to address diversity are allowed persist, especially when using an algorithmic approach [36]. When entities wish to be proactive regarding diversity, potential discrimination, or pursuing missed opportunities [78], they will need a path that passes muster against a range of challenges.

This paper thus seeks to offer techniques and legal analysis to enable companies to pursue legal and ethical hiring goals and face the question of *how to improve equal opportunity and employment practices without crossing into arguably illegal discriminatory practices.* The ideas discussed here are general and key takeaways can be applied to several stages in the hiring pipeline. That said, this paper uses the screening stage of employment to exemplify methods and analysis.

[1] As Primus has explained, "If equal protection requires the law to be thoroughly colorblind, then a statutory doctrine that requires racial classification and makes liability turn on the status of groups considered collectively is an equal protection problem" [86].

2 ALGORITHMS AND THE HIRING PROCESS

Many parts of the hiring process use algorithms as a way to manage and sort candidates—a practice which can be traced back at least 40 years [91, 92]. However, there are a number of junctures in the hiring pipeline at which bias can affect decisions. Job advertisements on various platforms can be targeted at specific audiences [30, 77]. Application rates can differ across groups due to presumed employer bias [83]. Data-driven tools for evaluating résumés can be biased due to inequalities in training data [62], imbalance in data [101], or differences in false positive/negative error rates in prediction algorithms leading to bias as a *downstream effect* [49]. Referral hiring can lead to favoritism [90]. Customer evaluations of freelancers can adversely impact certain groups [66]. Final hiring decisions can be influenced by human biases of the hiring committee [35]. After going through the hiring pipeline, candidates also see a significant difference in salaries offered [84], and retention rates can differ dependent on the work environment [44]. Indeed, societal biases are pervasive and can affect decisions made by experts [64].

In addition, when automated systems are used at any stage, missed opportunity (false negatives) with respect to minority candidates is often shrugged off as an artifact of the prediction model, necessary for overall accuracy [73]. These models often train on historic data, which can depict imbalanced selection rates across different groups of candidates, and these trends can be learned by automated methods [33, 39]. History can dictate future actions. In short, existing pipeline practices can reiterate and increase disparity in opportunity and outcomes. Although the hiring pipeline can be improved in many places, we find the screening stage to be particularly ripe for improvement, and we therefore focus the article on this stage for the reasons outlined below.

First, data-driven methods, by their nature, can pose a problem. Seemingly objective methods interact with real-world data, and so automated decisions can *reflect* and therefore, *reinforce* societal inequalities [38, 55]. Even when there is no intent to discriminate, and the decision system uses the same data and applies the same rule to all, there may be a disproportionate effect on a protected class (i.e., a group protected by law from discrimination, such as those defined by sex, race, age, etc.) [31, 32, 79]. The problems in screening map to the more general ones present when using data-driven decision-making in hiring. So, screening is a good lens through which to investigate the concerns around using algorithms and data in the employment context in general.

Second, algorithms are already used in screening. Such automated methods offer numerous advantages: speed, cost-effectiveness, potential objectivity, and uniformity in process. These properties may seem desirable at first glance from an ethical and fairness perspective; consistency in decisions is often a good thing, and a lack of human involvement would seem to minimize the role of implicit bias in hiring decisions [65]. Thus, automated methods have become commonplace in screening. Adjusting algorithms to address diversity or fairness concerns may be more palatable in an industry currently using automated processes than using algorithms in a heretofore un-automated process.

Third, changes at early stages of the hiring pipeline are vital to address later bias. Changes at later stages are only meaningful if they act on a fairly chosen pool of candidates. Without a fairly chosen candidate pool at those stages, efforts to address bias become empty theater. As such, we focus specifically on automated screening processes: *how should applicant-screening methods be developed?* These algorithmic tools should be designed to (a) select applicants of a desired quality, (b) satisfy some agreed upon fairness criteria, and (c) adhere to US anti-discrimination law.

3 BIASES IN DATA

We broadly refer to systematic inconsistencies in data which adversely affect certain groups as "bias." The first step in reducing discrimination is to understand the sources of bias. Unfair decisions can stem from many places, and identifying the origins of the bias allows for precise interventions. In the hiring process (automated or otherwise), applications will typically be assigned a score, thus allowing comparisons of applicants based on a single number or with respect to a single ranking of candidates [70, 89]. This evaluation metric can be hard-coded into an algorithm or developed dynamically, and in either case, can be unfair. A natural question is whether we can model this bias precisely and account for it within the algorithms to make them justifiably (provably) fairer.

Bias in evaluations can take different forms and be observed in different ways. For instance, a screening algorithm developed, *but not employed*, by Amazon penalized résumés which included the word "women's" due to data of past hiring trends in the company [45]. This algorithm penalized, for example, those who attended all-women colleges, and rewarded vocabulary typically used by men. In a similar vein, an empirical study showed that science faculty's assessment of résumés varied dependent on the gender of the student [84]. These are fairly blatant examples of discrimination, as toggling a protected attribute results in different treatment. Note that this form of unfairness—while blatant—can be hard to observe in practice, as applicants are never truly identical but for a small number of attributes.

Many cases of bias in evaluations, however, are more nuanced. Consider using SAT scores to screen candidates—a practice employers such as McKinsey, Bain, Goldman Sachs, and Amazon have been known to use even for candidates with advanced degrees [29, 48, 63]. Studies show that even if students are equally able to perform well on a test, if the test is announced to exhibit differences across groups, students in a negatively stereotyped group perform lower than the students in a non-stereotyped group [95]. Another study from 2013 shows that SAT scores are correlated with family income, potentially pointing to issues of access [50]. Inside Higher Education looked at SAT scores in 2015 and found that despite fee waivers and increased efforts to provide support and tutoring to low-income families, this correlation persisted across all three sections of the test. Compared to 2013, gaps in performance with respect to racial groups not only persisted but increased. This problem with SAT scores is further evident in a recent study by Faenza et al. [58], which showed a shift by approximately 200 points in SAT scores from schools with different *economic need indices*. Thus, an employer using SAT scores appears neutral but sets up a *pre-selected* pool.

These issues regarding bias in data raise important *design* questions for algorithmic intervention. When designing a decision-making algorithm, can we control for bias in historic data (thus avoiding Amazon's situation discussed above)? In other words, what steps can be taken to control for historic, economic, and/or social factors that are known to skew seemingly objective metrics such as the SAT?

4 APPROACHES TO ADDRESSING BIAS TO DATE AND THEIR LIMITS

A variety of algorithmic techniques have been proposed for coping with biased data and improving fairness, from pre-processing techniques which involve modifying data before feeding it to an algorithm [60]; to in-processing techniques, which modify the algorithm itself [72, 102]; to post-processing techniques, which modify decisions made by an algorithm after the fact [67, 71]. Current computer science literature highlights that merely scrubbing protected class information from an application may not help mitigate existing biases [46], and that algorithms have to use protected information to fix existing biases in data [54]. Using protected information, however, may put the hiring process at odds with anti-discrimination law. Other prevalent approaches include iteratively removing data which is correlated with protected information [102]; such approaches, however, may remove highly predictive information.

Algorithmic bias mitigation refers to the design of algorithms which perform well despite uncertainties about candidates' qualifications. This encompasses, for example, the design of procedures to select qualified candidates given biased data, or the design of algorithms which provably satisfy some notion of fairness. As discussed earlier, bias in evaluations can render bias-agnostic methods suboptimal [54, 56, 78, 88]; at the same time, imposing constraints such as demographic parity (i.e., proportional selection from different demographic groups) can hinder performance in some cases [43], which points to potential trade-offs between bias mitigation and quality of selections. In our approach, we will take the view of algorithmic bias mitigation given fine-tuned uncertainties in the evaluation of each individual.

Algorithmic Bias Mitigation. Attempts to mitigate bias often begin with an understanding of the nature of the bias, or in other words, the inconsistencies in measurement of the ability of candidates. Mitigating the impact of such inconsistencies is an instance-specific endeavor; no cure-all exists. Nonetheless, there is theoretical work on mitigating bias under various mathematical assumptions on bias and inconsistencies.[2] For instance, attempts have been made to address miscalibration of evaluations between multiple evaluators [98], and techniques have been developed for cases where some information is known about how biased each evaluator is in each evaluation [99]. Certain "coarse" sources of bias seem to be prevalent across demographic groups, and algorithms can be designed with these in mind. One might say these are the first

[2]This points to multiple issues in bias-mitigation. First, the assumptions on bias are difficult to justify empirically, as "ground truth" is seldom available (for example, the true ability of a candidate is never truly known, especially for candidates who are not hired). Second, it is difficult to assess bias-mitigation techniques for a similar reason: if one does not know the ground truth, then it is hard to quantify how good any decision is.

approximations to incorporate the knowledge of large trends visible broadly across demographic groups, such as are seen in SAT scores discussed earlier [58]. Addressing these coarser sources of bias from a theoretical point of view can provide insight in dealing with other forms of bias.

A recent mathematical model that captures the dependence of errors in testing over groups is the *group model* of bias. The model is based on the empirical work of Wennerås and Wold [100], and was introduced by Kleinberg and Raghavan in the context of offline selection (e.g., applicant-screening) [78], further studied by Salem and Gupta [88], Faenza et al. [58], and Blum and Stangl [37] in the context of selection problems. This model assumes that bias is fairly consistent within each demographic group, and thus evaluations offer accurate rankings within each group, but not across the groups. For example, once one accounts for difficulties in comparing one demographic group to another, there may be no way to confidently compare an 90% attained by a white male scholar Adam to a 85% attained by a Latina scholar Tia. But one can compare Adam against another white male scholar John with 83%, and note that Adam is better.

This model is at the same time appealing and dissatisfying in its simplicity. It is appealing in the sense that the model sheds light on best practices when the data is biased consistently for certain groups. That consistency indicates that information about group membership alone allows selection algorithms to reduce bias in selections. It is dissatisfying, however, in its coarseness, as it ignores intra-group differences in testing/evaluation errors and ignores any potential comparisons between groups. Adding to the example above, let us say that Tia also belongs to a low-income family, and we want to compare Tia to another Latina scholar May (not from a low-income family). This model does not account for such confounding variables of socio-economic status. Follow-up work by Celis et al. [40] proposed a multiplicative model of bias, wherein candidates in the intersection of biased groups face a consistently higher bias. This approach, however, again equalizes the amount of bias within each smallest "unique" group (e.g., male, white, and age above 45 or lesbian, Asian, aged 39). It may not be okay to equalize experience in these intersectional groups.[3] Indeed, whether a Chinese Asian, an Indian Asian, and a Filipino Asian faces the same amount of bias seems unlikely.

Current Industrial Practices. How then do companies actually hire candidates, while reconciling with anti-discrimination laws and biases in the hiring pipeline? In a recent survey, the only specific public claim made by vendors of pre-employment assessments was adherence to the 4/5ths rule—outlined in the 1978 Uniform Guidelines on Employee Selection Procedures—which requires that group-specific selection rates of any pre-screening are all within a factor of 4/5 of each other [87]. Yet this approach is coarse as it is agnostic to quality of candidates. Applying a 4/5ths rule in selection up front (e.g., as the current practice in the industry suggests [87]) does not change the perceived potential of candidates, nor account for uncertainties and biases in the data systematically. It can therefore simply set up the underrepresented group's candidates for failure, and lead to resentment and enlivening of negative stereotypes [59, 68].

The trade-offs in algorithmic approaches track legal issues. If an employer uses an algorithmic tool to evaluate and screen candidates, the employer may face legal challenges depending on the outputs of the tool. A likely challenge is that the tool created illegal disparate impact. Disparate impact addresses when "facially neutral policies or practices have a disproportionate adverse effect or impact on a protected class" [28, 61]. The disparate impact doctrine is thus supposed to address situations where intent is not at hand or cannot be ascertained [96]. In short, outcomes based on unaware algorithms may fit quite well with disparate impact challenges, because unaware algorithms are facially neutral, may lack intent to discriminate, and nonetheless yield statistically discriminatory results.[4]

The prospect of a disparate impact claim may lead an employer to design an algorithm that takes protected class status into account. However, this approach may run into a disparate treatment challenge. Disparate treatment is the legal doctrine that prohibits intentional use of race or other protected classes in making an employment decision. Thus, we return to the paradox described above, because it seems that an employer is trapped between using facially neutral systems that reflect systemic and historically conditioned, biased results or facing lawsuits for using aware systems to mitigate such effects. This paradox is exacerbated by current legal scholarship debating what algorithmic interventions to address bias, if any, are allowed and the implications of the lawsuit *Ricci v. DeStefano*, in which an action by the City of New Haven that tried to account for disparate impact of an administered promotion test led to litigation that was decided against the city. In Section 5, we outline the *poset approach*, which we argue provides a way to solve the hiring paradox. Section 6 turns to an in-depth discussion of *Ricci v. DeStefano* and explains how the poset approach fits within legal rules so that one can use a bias aware approach and yet maintain individualized assessments of candidates.

5 A NEW APPROACH: COPING WITH UNCERTAINTY USING PARTIAL ORDERS

As discussed in Section 3, coping with uncertainties in data is a fundamental problem in applicant screening systems, as well as in data-driven decision-making more generally. Here, we will discuss one method, the *poset approach*, for applicant-screening in the face of uncertainty which has emerged recently in the computer science literature [88]. In Section 6, we will use this approach as a vehicle for discussing the legality of algorithmic bias mitigation in hiring.

Consider the following scenario: there are three candidates A, B, and C, with ability scores of 82, 68, and 67, respectively, and you wish to grant interviews to two of them. The ability scores are known to be a strong predictor of job performance, but are only known to be accurate up to 3 points. In this case, there is a significant chance that C is a better candidate than B, but the utilitarian approach of selecting the highest-scoring candidates

[3] One underlying problem with this model is the assumption of group membership, which may not even be accurate in practice.

[4] Despite the fact that the 4/5 rule is mentioned as evidence of disparate impact in the 1978 Uniform Guidelines on Employee Selection Procedures, there is no precise quantification of disparate impact. The 4/5 rule is often used as a trigger for litigation, but other statistical tests have been used in courts as well [82, 94].

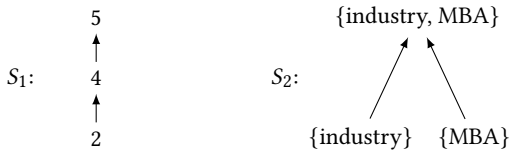

Figure 1: Hasse diagrams for posets S_1 and S_2 described in Section 5.

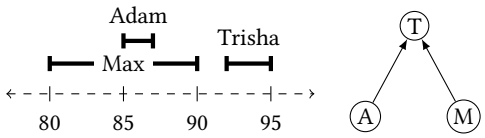

Figure 2: Score ranges and resulting Hasse diagram for the scenario in Example 5.1.

would routinely select A and B. The core idea behind the poset approach is that the latter approach is unfair to C, or more generally, that ignoring uncertainty can result in unfair decisions. In other words:

> *Some applicants, due to insufficient or inaccurate data, cannot be reliably ranked. The solution need not involve producing a (possibly inaccurate) ranking. Instead, allowing for partial rankings can open the door to fairer decisions.*

The poset approach, which we explain in more detail below, makes use of a mathematical structure called a *partially ordered set*, or *poset*, which can be used to encode uncertainty in ordinal information. Consider, for example, a set $S_1 = \{4, 2, 5\}$ of true hirability of three candidates (which is often not observable in practice). This set is called *totally ordered* since any pair of the scores can be ordered (i.e., ranked) with respect to the relation $\leq$. In other words, we can rank the scores: $2 \leq 4 \leq 5$, thus inducing an order on the candidates.

However, in practice, one cannot observe directly how good a candidate might be at their job. This is where partial orders can help us. Intuitively speaking, one can think of a partial order as a set of comparisons, which may not cover all pairs of candidates (i.e., a total order with some comparisons missing). For example, consider a candidate D who has experience in industry, a candidate E who has experience in industry *and* who has an MBA, and a candidate F who has an MBA. Considering these traits as binary (yes/no) attributes, one can represent their qualifications as the set $S_2 = \big\{\{\text{industry}\}, \{\text{MBA}\}, \{\text{industry}, \text{MBA}\}\big\}$. From the given information, one might rank E above both D and F, since E is qualified with respect to both measures, and the other candidates are only qualified with respect to one. However, D and F might be considered *incomparable*, since their qualifications are complementary. In this case, S_2 is a partially ordered set,[5] but not a totally ordered set. A poset is often visually depicted using its *Hasse diagram*, which is a directed graph in which edges represent orderings. For example, the Hasse diagrams for S_1 and S_2 are shown in Figure 1.

The *poset approach* is the process of (1) forming a partial ranking (i.e., a partial order) of the candidate pool based on uncertainties, inaccuracies, or biases in data, and (2) making selections based on this poset. By making selections in this way, one can concretely take uncertainty into account and, say, avoid routinely harming candidate C in the example above. This approach can mitigate biases in the evaluation metric; e.g., if a group is underrepresented in training data and experiences large errors in the resulting ML model, the poset approach can confer benefit of the doubt to those underrepresented candidates. We next illustrate how posets can model uncertainty using two examples.

Example 5.1. The poset approach can illustrate how one can account for uncertainties while also avoiding prohibited discrimination based on gender.[6] Using the poset approach, one may incorporate demographic context of the candidates and quantify uncertainy in their evaluations (either by directly observing the context, or by unsupervised methods such as clustering). Suppose that in a training dataset, nonbinary candidates are underrepresented, and as a consequence have high variance in errors in the prediction model. One may find that a nonbinary candidate Max has a wide score range of 80-90% (e.g., due lack of training data on nonbinary candidates), another male candidate Adam has a score between 85-87%, and a third female candidate Trisha has a score between 92-95% (see Fig. 2). Now, using only the score ranges to compare candidates, Trisha compares favorably to Max, but it is unclear if Max is more qualified than Adam as their ranges overlap. In this case, we can think of Max and Adam as mutually incomparable. The poset approach therefore allows for individualized treatment of inconsistencies in data processed.

Example 5.2. Suppose that three candidates are to be selected based on two attributes: work experience and college GPA. You have set cutoffs for each of these attributes and only wish to select candidates exceeding each cutoff. See Figure 3 for a depiction of the candidate pool, where each color represents a particular demographic group. Let the colored areas around each candidate node represent a "confidence region;" i.e., with some high degree of confidence, the candidate's latent ability lies in the drawn region. Note that we can infer partial rankings from these confidence regions in a similar way to Example 5.1: if the confidence region of candidate A is strictly above and to the right of the confidence region of candidate B, then A is ranked above B.

Using only raw scores, only the two blue candidates meet the cutoffs. However, taking confidence regions into account, we see that the two green candidates might meet the cutoffs as well. How, then, should one choose three candidates among the green and blue ones? One way to do so is to look at the partial ranking induced by the confidence regions (shown using arrows in Figure 3). In this partial ranking, there are three candidates who are maximally ranked (i.e., are not ranked below any other candidates): the two blue candidates and the right-most green candidate. This is one possible justification for selecting these candidates.

[5] A relation $\preceq$ is a partial order on a set S if three conditions hold for all $a, b, c \in S$: (1) $a \preceq a$; (2) if $a \preceq b$ and $b \preceq c$, then $a \preceq c$; and (3) if $a \preceq b$ and $b \preceq a$, then $a = b$. One can check that all these properties are satisfies for the set S_2.

[6] As recently as June 15, 2020, the Supreme Court of the United States ruled that the Title VII of the Civil Rights Act prohibits discrimination on the basis of sexual orientation and gender identity [27, 97].

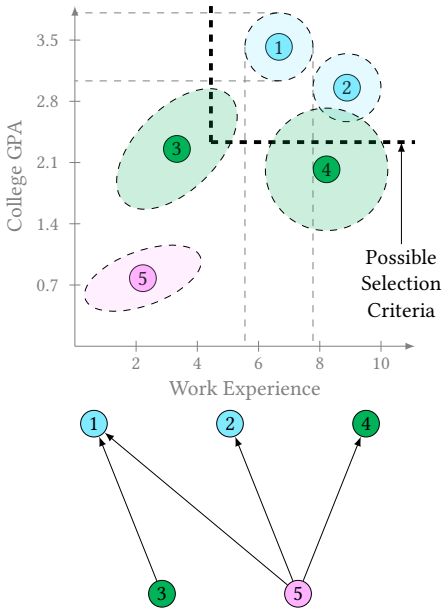

Figure 3: A depiction of the work experience and college GPA of five candidates coming from three groups (differentiated by color). The thick dashed line represents a possible selection criteria which, in this case, imposes a threshold on each of the two attributes. Confidence regions are drawn around each data point to indicate, say, 95% confidence in the inclusion of a candidate's true ability. Given these confidence regions, one can construct a partial ranking, as depicted by the Hasse diagram on the bottom. Arrows between candidates indicate ranking with certainty with respect to both attributes (e.g., Candidate 1 is ranked higher than Candidate 5 since the best work experience score (≈ 4) in the confidence region of Candidate 5 is worse than the worst work experience score (≈ 5.5) of Candidate 1, and similarly for College GPA. Note that there may be other reasonable ways of constructing partial rankings as well.

The process outlined in these examples (forming score ranges/regions for each candidate and inferring comparisons therefrom) can be applied quite generally, and allows for explicit treatment of bias in data. Data-driven techniques, such as estimating latent group-bias in a machine learning model, can be applied to generate these score ranges, which in turn can induce a partial ranking. Such methods can be used to avoid penalizing applicants who come from underrepresented groups, who are more likely to face inaccurate evaluation via machine learning models. A recent paper by Emelianov et al. shows that groups with high error variances can receive worse treatment, even if the evaluations are unbiased for all candidates [56], pointing to the need for interventions like the poset approach that take uncertainty into account.

The poset approach in practice. We end this section by providing a framework for using the poset approach in practice. While this framework does not encompass every possible use of the poset approach [88], it will describe the process from beginning to end and put the poset approach in broader context.

Step 1. Clean-up and Process Past Hiring Data. To start, collect data from previous hiring cycles. This data might include scores derived from textual analysis of résumés, test scores for job-related tasks (e.g., computer programming test scores), automated scores based on analysis of video interviews [87], college GPAs, courses taken, years of work experience, job performance of those who were hired, and so on.

Step 2. Quantify Uncertainty and Bias. Use data analysis to quantify potential data biases. Clusterings, for example, can help determine if evaluations unfairly favor one group over another. Looking at the data along different demographics (e.g., based on race, gender, age) can point to potentially discriminatory decisions in the past. Use social science studies (e.g., [95]) that highlight the impact of social status on the considered metrics (e.g., standardized test scores) to guide analysis. This will help highlight qualitative and quantitative reasons for disparities in the past hiring data.

Step 3. Construct a Partial Order. Trends identified in Step 2 can be used to construct a partial ranking of candidates. For example, score ranges can be constructed for each attribute of interest using a prediction model and estimates of its error variances, and these ranges can inform partial rankings as in Figure 3. These ranges can take into account distributional differences across protected attributes, differing error variances due to training data imbalance,[7] observed inaccuracies in past predictions, and so on. Predicted scores can be modified to be distributionally similar to the true scores on a group-by-group basis, and score ranges can be constructed around these transformed scores. One can construct a partial order to account for group-specific errors even if the evaluation metric was provided by a third party and its inner workings are unknown to the user. Unsupervised methods such as clustering can be used without knowledge of protected information, and a partial order can be constructed based on the extent of uncertainty or bias in each cluster. These approaches are discussed in more detail in Appendix A, and more examples are given in [88].

Step 4. Adapt Selection Algorithms. Once the partial ranking has been constructed, selections need to be made. Presumably, a hiring committee already has a screening process (automated or otherwise) which aligns with the goals of the employer. In order to implement the poset approach, this screening process must be adapted to take a partial ranking as input instead of numeric scores or a total ranking. Typically, this can be done by prioritizing maximality and randomizing wherever incomparabilities necessitate (see [88] for an example of this in an online setting).

Step 5. Auditing for Policy Compliance. The entire hiring pipeline may be subject to auditing for compliance with anti-discrimination policy. It is prudent to document and be able to justify each decision made in the hiring process, particularly those pertaining to the four steps outlined above. For example, one should be able to explain

[7] This refers to the observation that a group which is underrepresented in training data often experiences large errors in a resulting prediction model. In the poset approach, these larger errors could translate to larger score ranges for the underrepresented group. Note that the groups in question could come from a clustering and need not be demographic groups.

how the partial ranking was constructed and be able to justify those decisions by pointing to data and relevant research. A deeper discussion of the legality of the poset approach (and algorithmic bias mitigation more generally) is in Section 6.

6 DISCUSSION AND BEST PRACTICES—LAW, MATHEMATICS, AND POSETS IN PRACTICE

We now return to the business cases with which we started and the tensions they present regarding diversity, equity, and legal interests. On the one hand, firms are seeking to address diversity regardless of a history of discrimination. On the other hand, when evidence of past or present practices creating barriers is found, companies addressing those practices are pursuing affirmative action plans. In general, a firm that does little to account for race, gender, and other protected classes may find it has created disparate impact; and yet, when that firm seeks to take protected classes into account, such steps may violate the ban on disparate treatment.

We offer that in the unlikely case where a firm has no reason to believe that norms, traditions, or societal inequalities are negatively affecting the ability for members of a protected group to pass through the stages of the hiring pipeline, action may be possible under diversity interests but not required by law. At least two shifts point to increased diversity activity. First, many companies have made public commitment to large steps to address diversity in employment. Second, there is a new push for companies to disclose workforce diversity data, which has resulted in 82 of the top 100 companies doing so. The public imperative combined with the data supports companies taking the initiative to address workforce imbalances regardless of legal requirements to do so [74]. In contrast, as a matter of affirmative action, a firm with evidence of discrimination seeking to address imbalances in its workforce should be able to take steps to do so. Such steps could involve, for instance, scoring applicants using a machine learning model and developing confidence intervals around scores using the poset approach. From a legal perspective, it is important to be able to support the legality of each action, from the decision to address diversity to the decision to use protected class information, to each design choice in the algorithm, to each adjustment to future rounds of hiring.

The beauty of the poset approach is that it is agnostic to the motivation (diversity or addressing discrimination via affirmative action) behind a company's plan. To be clear, whether a purely diversity-driven plan is legal is an unsettled question and beyond the scope of this paper [36, 57]. Nonetheless, because of the current drive to address inequity, we expect this question to arise in the near future and suggest that the poset approach would aid and support such efforts. Furthermore, because many announced diversity programs are likely backed by data about imbalances and unnecessary barriers to employment, such efforts will likely be seen as affirmative action plans under the law. Thus, in this section, we address the core question of how well the poset approach stands up to legal scrutiny as an allowed method to address affirmative actions plans.

Given that efforts to modify evaluation mechanisms or selection algorithms can raise both disparate impact and disparate treatment issues, we now use a hypothetical employer perspective in line with Microsoft's and other companies' announced goals to suggest best practices. Insights are derived from a series of questions about how to identify workforce imbalances (Section 6.1) and how to address said imbalances (Section 6.2).

6.1 Diagnosis

Q1: *An employer is concerned that its workforce under-represents women and minorities. May they do anything to change their current hiring practices?*

Yes. The purpose behind Title VII is "[T]o achieve equality of employment opportunities," and Congress "directed the thrust of the Act to the consequences of employment practices, not simply the motivation" [2]. That means that "unnecessary barriers to employment" must fall, even if "neutral on their face" and "neutral in terms of intent" [1]. Federal courts have disallowed a host of hiring and promotion practices that "operate[d] as 'built in headwinds' for minority groups" [19]. In addition, the Supreme Court has upheld the legality of employment plans to address discrimination without reference to its past practices or evidence of a possible violation of the law [8].

To take action, an employer "need[s] to point only to a 'conspicuous ... imbalance in traditionally segregated job categories' " [8]. Logically, this requirement implies that initial, proactive analysis identifying the imbalance problems can serve as justification for adjustments to hiring practices. As such, employers can and should use data science and analytics to identify the imbalance in their hiring pipeline that it seeks to address [69, 75, 80].

As one example, the employer can use human resources data to examine its employment practices. First, it can audit its current workforce and get fine-grained information about who works at the company and at what levels. Such an approach allows the company to look beyond simple questions such as "Does it have an equal number of men and women in the workforce?" Instead, the company can see the gender and minority makeup at different levels of employment such as upper management, upper-middle management, middle management, administration, hourly workers, contractors, and so on. Visualizing the data with pie-charts or heat maps will provide clear, vivid ways to see the current state of affairs. Second, after such a study, the company can see potential sources of issues. It may find that women and minorities rarely move beyond middle management, are rarely interviewed for promotion, or that screening to date has not selected, or under-selected, women and minorities for interviews to be potential employees. At a general level, these types of analyses support the case that there is something to fix. This gets us to the next step in the process.

Q2: *If a company finds that women and minorities are rarely interviewed and further finds that screening to date has not selected, or under-selected, women and minorities for interviews to be potential employees, do these conditions support allowing an employer to use protected-class information to build or apply a bias-aware algorithm at the screening stage?*

Identifying a problem with a screening process or a structural problem in the company's workforce reveals a clear "unnecessary barrier to employment" even if the algorithm is neutral on its face and in intent. For example, if men tend to be scored higher than

women (e.g., as in Fig. 4), then a facially neutral selection algorithm would disproportionately select men, even if true ability is similar across genders. In general, the identified, strong evidence of bias in current algorithmic sorting in the hiring process, including the screening stage, should constitute the sort of "built in headwind[] for minority groups" that the law seeks to eliminate. With sufficient evidence of bias and systemic barriers to equality of employment opportunities, an employer can make a case for using bias-aware algorithms.

6.2 Corrective action

Voluntary action to comply with the goals of Title VII is not only allowed; it is favored [9]. Nonetheless, in some cases, trying to further the goals of Title VII to address discrimination raises the paradox where one approach looks like disparate impact and a corrective action looks like disparate treatment. What can a company actually do?

Q3: *May an employer use protected-class information to increase diversity among interviewees?*

This question is complex as it entwines various parts of the process that need to be slowly unpacked. A recent case *Ricci v. DeStefano* [20] illustrates some problems and provides guidance on allowed and prohibited actions.

Background. In *Ricci v. DeStefano*, the City of New Haven had developed a test for firefighter promotion with the help and validation of experts. When administered, 77 people took the lieutenant exam: "43 whites, 19 blacks, and 15 Hispanics. Of those, 34 candidates passed: 25 whites, 6 blacks, and 3 Hispanics." 41 people took the captain's exam: "25 whites, 8 blacks, and 8 Hispanics. Of those, 22 candidates passed: 16 whites, 3 blacks, and 3 Hispanics." Despite the experts' opinions and validations of the test, the City rejected the results because the pass rate caused the city to believe it might be sued for disparate impact. The Supreme Court did not allow this after-the-fact change, because New Haven's actions relied on race (the race of those who passed the test) to reject the results, and in that sense, New Haven engaged in disparate treatment. Thus, it may appear that an entity cannot account for and alter employment practices when there is evidence of potential disparate impact in the entity's practices, because such changes will necessarily constitute disparate treatment [33]. That is incorrect [76].

Analysis. As the Supreme Court put it, not allowing an entity to account for race to avoid disparate impact liability "if the employer knows its practice violates the disparate-impact provision," is contrary to "Congress's intent that "voluntary compliance" be "the preferred means of achieving the objectives of Title VII" [24]. This rule, however, does not mean an entity can simply assert there has been a history of past discrimination and so a need to throw out a practice, because that might lead to "an unyielding racial quota" [25]. As stated above, the entity has to show why the change is needed in light of the goals of Title VII. In addition, the timing of when an entity makes changes matters.

The way the test was developed and administered by New Haven doomed the City's decision to reject the test's outcomes. New Haven began well by hiring experts to design a likely *valid test*. The City spent $100,000 on experts on designing the tests for fire departments [21]. The experts conducted interviews, went on ride-alongs, interviewed incumbents at the promotional level at issue, and designed "job-analysis questionnaires and administered them to most of the incumbent battalion chiefs, captains, and lieutenants in the Department" [21]. As the Supreme Court noted, "At every stage of the job analyses, IOS [the company that developed the test], by deliberate choice, oversampled minority firefighters to ensure that the results—which IOS would use to develop the examinations—would not unintentionally favor white candidates" [22]. Once the test was approved, New Haven set a 3-month study period and gave candidates a study guide including the "source material for the questions, including the specific chapters from which the questions were taken" [22]. Nonetheless, after the tests were given, the results indicated disparate impact [23].

The city's ex-post actions were the problem. The Court rejected "invalidating the test results" after the fact without "a strong basis in evidence of an impermissible disparate impact" [26]. The ex-post rejection of the results created "visible victims"—that is, those who studied for the test, passed, and whose hard work was discarded [86]. After the city gave the test, it needed strong evidence that the test would be invalidated if the city were sued for disparate impact and lose, because otherwise those who had passed would be harmed. The Court did not see such evidence and so did not allow the city to reject the results.

Answer to Q3. Designing a screening system is quite different than what happened in *Ricci*. *Ricci* was about a later stage of employment (i.e., promotions), and it involved a test for which many test-takers had prepared, including spending money on test preparation aid. The advantage of building a screening system is that the actions are ex-ante, and the system is not a test for which someone can prepare [36]. Unlike in *Ricci*, where applicants were seen as having an expectation that a potentially valid test for which they could study be accepted, designing and using a screening algorithm occurs at an earlier stage of the hiring process where no hiring or promotion decision is made. Thus in designing a screening algorithm, one might observe selections over time and change the parameters to create a more representative sample of qualified candidates, including making adjustments during the "training" of the algorithm. These steps are analogous to the design steps—such as making overt choices and oversampling at every stage to ensure that the test did "not unintentionally favor white candidates"—taken by New Haven and of which the Supreme Court wrote with approval [22]. In other words, designing and vetting a screening system to ensure that the results are not having discriminatory outcomes should be legal.

Recall that one of the goals of Title VII is to reduce, if not eliminate, "unnecessary barriers to employment." The *Ricci* Court did not "question an employer's affirmative efforts to ensure that all groups have a fair opportunity" at a given stage of the hiring process. An employer is allowed to examine "how to design. . .[a] practice in order to provide a fair opportunity for all individuals, regardless of their race" before deploying it [26]. Designing a screening algorithm is by its nature an ex-ante event for which a candidate cannot prepare in the way one might for a test.

In short, if Question 2's requirement is met, an employer should be able to develop a bias-aware algorithm to avoid disparate impact. Of course, we still need to address the validity of the new practice and what is allowed in its design, which brings us to the next question, which we partially answer through the lens of the poset approach.

Q4: *What is allowed in the design of a bias-aware algorithm? Can it be designed to improve the yield of whom to interview?*

This is one of the grand challenges in this area. Let us focus our attention to the proposed poset approach, and draw arguments from the Supreme Court's decision in *Johnson*. The key to using a bias-aware algorithm such as the poset approach of Salem and Gupta is to establish the facts and evidence of a need to address bias (or more generally, inconsistencies in the data) as set forth above, and then to build a plan that assesses individuals rather than setting up a purely number-driven process with quotas for each category [9]. If a plan is "blind hiring," that is, dictates hiring "solely by reference to statistics" or "by reflexive adherence to a numerical standard," the plan is not likely to be allowed [11]. But, if a plan takes "numerous factors. . .into account in making hiring decisions, including specifically the qualifications of [all] applicants for particular jobs," the plan may take a protected class into account as part of the overall evaluation [9]. In that sense, the protected class status "may be deemed a 'plus' in a particular applicant's file, yet it does not insulate the individual from comparison with all other candidates for the available seats" [3, 12].

Comparison does not require pure, numeric ranking; indeed, that might tip into the sort of "blind hiring" that is disfavored. As the Sixth Circuit stated, the "practice of rank-order hiring from a single list grouping together males and females was impermissible under Title VII because the City could not establish that higher scores on the test meant better job performance." [15]. The Second Circuit has explained that evaluations should be sufficiently correlated with job performance to induce a rank ordering, where the quantification of "sufficiently correlated" may depend on the extent of adverse impact of the evaluation metric [5]. The Sixth Circuit additionally asserted that a certain cognitive ability test could not be used as the sole basis for a rank-ordering despite being predictive of job performance, since the test failed to measure certain qualities of interest. Rank orderings based on evaluations should therefore not be thought of as implicit to a screening practice, but instead as a design choice which must be justified [15].

Discretion in comparison of candidates is allowed when it is part of the overall, individual assessment. For example, in *Johnson v. Transportation Agency of Santa Clara County*, two candidates were deemed well-qualified based on a range of metrics, such as experience, background, and test scores taken together. But each candidate had differences within a given metric. One had more clerical work and more road maintenance work; the other had more experience at a specific part of the business. As for test scores, the man scored 75 on the interview portion of the assessment and the woman scored 73. The employer had set 70 as the minimum threshold for the interview and seven applicants crossed the 70 mark. The range of acceptable scores was 70 to 80 [7]. The woman was given the promotion over the man who had the higher score. Because the scores were within the range of acceptable scores and the final hiring manager looked at a set of metrics with gender as "but one of numerous factors he took into account in arriving at his decision," the plan's incorporation of bias-awareness, here regarding gender, was allowed [12].

Other cases also acknowledge the need for an approach beyond using an absolute score or ranking. Given problems with rank-ordering, the Second Circuit of Appeals has allowed a rather coarse approach where an employer may "acknowledge his inability to justify rank-ordering and resort to random selection from within either the entire group that achieves a properly determined passing score, or some segment of the passing group shown to be appropriate" [6]. Courts have also indicated an acceptance for more nuanced methods. For example, the act of "banding," or considering score ranges instead of singular scores, has been accepted to account for inaccuracies in evaluation [16, 17]. Although these cases consider banding in a quite limited sense in that scores ranges are centered on original scores and are of uniform length, they support that one might relax the assumption of an absolute ranking of candidates.

In language that tracks the poset approach, the Second Circuit has also acknowledged "that small differences between the scores of candidates indicate very little about the candidates' relative merit and fitness" [6]. Thus the court embraced an approach that assessed "a statistical computation of the likely error of measurement inherent" in its exam. The employer then used that measurement to set up zones of candidates clustered by test scores within that error measurement. That practice was seen as a good solution to "insur[e] compliance" with Title VII. The Second Circuit explained, "by creating a more valid method to assess the significance of test scores, [the approach] eliminated the central cause of the adverse impact, i.e., the rank-ordering system, while assuring appointments on the basis of merit." As such, if one is able to use protected information (as in *Johnson*, or in the context of a valid affirmative action plan [36]), then the banding cases provide guideposts for adopting the poset approach as described in Section 5.

Answer to Q4. 1. An algorithmic approach should be allowed. A takeaway from *Johnson* and the cases on banding and rank-ordering is that a precise numerical score is not necessarily indicative of an applicant's potential, and courts welcome approaches that better compare candidates. Thus, score ranges can be used as part of an applicant-screening procedure. This supports the use of score ranges to account for uncertainties in evaluations, as outlined in Section 3.

Further, note that incorporating the poset model of bias is not the same thing as normalizing distributions of scores across groups. When we normalize scores across groups, we are essentially transforming all scores so that group-specific distributions look similar, and this process results in a full ranking of applicants. In contrast, the poset approach intentionally does not reduce each applicant to a number and allows for incomparabilities between applicants. This allows for a more individual treatment of candidates, where uncertainty in rankings can be acknowledged. The result is that applicants are assessed as individuals, potentially in a more mathematically sound way.

2. There are rules about when bias-aware algorithms can be used. Recall that the stage at which an entity uses bias-aware algorithms matters. In the promotion context of *Johnson*, the Court gave a further reason the plan was allowed. Unlike *Ricci*, where applicants were seen as having an expectation that a potentially valid test for which they could study be accepted, there was "no absolute entitlement" to the position at issue in *Johnson*. The entity had seven qualified and eligible applicants, and choosing one over the other "unsettled no legitimate, firmly rooted expectation" of any of the candidates. By extension, a bias-aware applicant-screening plan that used a protected class as part of an overall assessment then had all selected applicants compete on the same metrics should be allowed under the law.

3. There are legal rules on the goals of any hiring plan. The law respects plans that seek to remedy an imbalance and that do not set aside positions for a given group while also conducting annual reviews of goals as it fashions future rounds of hiring and promotion [14]. One may work "to attain a balanced work force, not to maintain one" [13].

The *Johnson* Court also noted with approval that "the Plan sought annually to develop even more refined measures of the underrepresentation in each job category that required attention" [10]. This idea of not maintaining a balanced workforce reflects the idea that an entity cannot use a plan that sets up quotas to maintain balance based purely on class statuses. By extension, suppose balance is achieved in a company through bias-aware methods, and they notice this by continuous monitoring of their hiring practices (in a sense, returning to Question 1). The company may then have to stop using bias-aware methods, even if demographic imbalance persists in the general workforce for that line of work.

4. The poset approach does not impose quotas. In contrast to methods described in some recent work [56], using score ranges (e.g., using the poset approach) instead of raw scores does not set up a quota system.[8] When using the poset approach, selection rates may be influenced by protected information (e.g., when accounting for observed, group-specific biases), but such protected information is not necessarily a determining factor in selection decisions. For example, the poset approach could result in a set of candidates which is less demographically proportional than what raw scores might produce (see, e.g., Figures 6-7 in Appendix B), or more demographically proportional (see, e.g., Figures 4-5), depending on the data and the ascertained uncertainty therein.

7 CONCLUSION

We summarized recent work in the context of hiring, with a focus on screening algorithms. We highlighted the seeming paradox of mathematics, law and practice that a company might observe workforce imbalance due to its past practices, but the solutions to correct for this imbalance are either at a contradiction with mathematics or anti-discrimination law. The new poset-based approach [88] provides a framework for incorporating uncertainties in rankings into a candidate-screening practice which allows, for example, hiring committees to base decisions on confidence intervals of ability scores. This approach can potentially be legally justified based on past disparate impact and can be adjusted over time as the data grows and hiring goals evolve; and thus can help avoid having a static plan as the law requires.

No approach, however, is a fix-all solution. The poset approach cannot discount for undetectable errors or modeling errors due to missing data. The lengths of the intervals impact the quality of selections. Further, two different mathematical approaches could be used to define score ranges for candidates and result in different sets of selected candidates. A legal dispute may require addressing which one of these approaches is more valid. Further, there is an "are we there yet?" issue built into the Supreme Court's rulings. That is, it may be unclear at which point a workforce becomes "balanced" and the current plan must be replaced. Although the poset approach is adaptive, detecting where there is no longer any impact of societal biases in the data is non-trivial, and we leave this as an open question.

For any intervention in an existing framework, one has to consider if the intervention is serving those for whom it is designed [42]. Partially ordered sets that are interval-based might create an impression that certain underrepresented minorities carry high uncertainty in their potentials and as a result, lead a risk-averse hiring committee to reject those candidates. On the contrary, the poset approach is able to highlight *missed opportunities* in representation in the hiring pipeline. Taking uncertainties into account can expand and improve the talent pool to include candidates who are qualified and would have been competitive had there been no bias in the data. Thus, we believe that the analysis presented here can pave the way forward for hiring qualified candidates in a fair way in the evolving legal landscape.

ACKNOWLEDGMENTS

The authors thank Jason Bent, Justin Biddle, Kimberly Houser, Pauline Kim, Orly Lobel, and the participants of the Data, Law, and Ethics Virtual Conference hosted by the University of Indiana, Kelly School of Business, as well as the participants of the Privacy Law Scholars Conference 2021 for their helpful comments on an earlier draft of this work.

[8] One can set aside seats for interviews as happens with the Rooney Rule in the NFL, but such a rule is best-protected by following the legal constraints for affirmative action plans. See, e.g., https://www.aclusocal.org/en/inclusion-targets-whats-legal. Further, quota-based approaches at screening stages may create a pool of candidates destined for later rejection. The poset approach enables selections based on the possibility of a candidate being qualified and so better fits legal requirements at any stage of hiring.

REFERENCES

[1] 1971. Griggs v. Duke Power Co., 401 U.S. 424, 431.
[2] 1971. Griggs v. Duke Power Co., 401 U.S. 424, 432.
[3] 1978. Regents of University of California v. Bakke, 438 U.S. 265, 317.
[4] 1979. EEOC Guidelines on Affirmative Action, 29 C.F.R. § 1608.1(c).
[5] 1980. Guardians Ass'n of New York City v. Civil Serv, 630 F.2d 79 (2d Cir.).
[6] 1983. Kirkland v. N.Y. State Dep't of Correctional Serv., 711 F.2d 1117, 1133 (2d Cir.).
[7] 1987. Johnson v. Transportation Agency, Santa Clara Cty., 480 U.S. 616, 623-624.
[8] 1987. Johnson v. Transportation Agency, Santa Clara Cty., 480 U.S. 616, 627.
[9] 1987. Johnson v. Transportation Agency, Santa Clara Cty., 480 U.S. 616, 631.
[10] 1987. Johnson v. Transportation Agency, Santa Clara Cty., 480 U.S. 616, 635.
[11] 1987. Johnson v. Transportation Agency, Santa Clara Cty., 480 U.S. 616, 636-637.
[12] 1987. Johnson v. Transportation Agency, Santa Clara Cty., 480 U.S. 616, 638.
[13] 1987. Johnson v. Transportation Agency, Santa Clara Cty., 480 U.S. 616, 639.
[14] 1987. Johnson v. Transportation Agency, Santa Clara Cty., 480 U.S. 616, 640-641.
[15] 1995. Brunet v. City of Columbus, Ohio, 58 F.3d 251, 255 (6th Cir.).
[16] 1998. Boston Police Superior Officers Fed'n v. City of Boston, 147 F.3d 13.
[17] 2005. Bradley v. City of Lynn, 403 F. Supp. 2d 161.
[18] 2006. Title VII, 29 CFR Parts 1600, 1607, 1608.
[19] 2009. Ricci v. DeStefano, 557 U.S. 557, 632.
[20] 2009. Ricci v. DeStefano, 557 U.S. 557.
[21] 2009. Ricci v. DeStefano, 557 U.S. 557, 564.
[22] 2009. Ricci v. DeStefano, 557 U.S. 557, 565.
[23] 2009. Ricci v. DeStefano, 557 U.S. 557, 567.
[24] 2009. Ricci v. DeStefano, 557 U.S. 557, 580-581.
[25] 2009. Ricci v. DeStefano, 557 U.S. 557, 583.
[26] 2009. Ricci v. DeStefano, 557 U.S. 557, 585.
[27] 2020. Bostock v. Clayton County, 590 U.S. ___.
[28] 42 U.S.C. § 2000e-2(k)(1)(A).
[29] Accessed January 17, 2022. McKinsey's online application FAQs: Careers. https://www.mckinsey.com/careers/application-faq
[30] Julia Angwin, Noam Scheiber, and Ariana Tobin. 2017. Dozens of companies are using Facebook to exclude older workers from job ads. *ProPublica, December* (2017).
[31] Solon Barocas. 2014. Data mining and the discourse on discrimination. In *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining*. 1–4.
[32] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.
[33] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[34] Dina Bass and Josh Eidelson. 2020. Microsoft plan to add Black executives draws U.S. Labor inquiry. *Seattle Times* (October 6, 2020). https://www.seattletimes.com/business/microsoft/microsoft-plan-to-add-black-executives-draws-u-s-labor-department-inquiry/
[35] Ashley B Batastini, Angelea D Bolaños, Robert D Morgan, and Sean M Mitchell. 2017. Bias in hiring applicants with mental illness and criminal justice involvement: A follow-up study with employers. *Criminal Justice and Behavior* 44, 6 (2017), 777–795.
[36] Jason R Bent. 2019. Is algorithmic affirmative action legal. *Geo. LJ* 108 (2019), 803.
[37] Avrim Blum and Kevin Stangl. 2020. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?. In *Symposium on Foundations of Responsible Computing (FORC)*, Vol. 1.
[38] Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias. (2018).
[39] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[40] L Elisa Celis, Anay Mehrotra, and Nisheeth K Vishnoi. 2020. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 369–380.
[41] Ronald Christensen. 1996. *Analysis of variance, design, and regression: applied statistical methods*. Page 173. CRC Press.
[42] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).
[43] 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.
[44] Vedant Das Swain, Koustuv Saha, Manikanta D Reddy, Hemang Rajvanshy, Gregory D Abowd, and Munmun De Choudhury. 2020. Modeling organizational culture with workplace experiences shared on glassdoor. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[45] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters* (Oct 2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
[46] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.
[47] Deven R Desai and Joshua A Kroll. 2017. Trust but verify: A guide to algorithms and the law. *Harv. JL & Tech.* 31 (2017), 1.
[48] Shalia Dewan. 2014. How Businesses Use Your SATs. *New York Times* (Mar 2014). https://www.nytimes.com/2014/03/30/sunday-review/how-businesses-use-your-sats.html
[49] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.
[50] Ezekiel Dixon-Roman, Howard Everson, and John Mcardle. 2013. Race, Poverty and SAT Scores: Modeling the Influences of Family Income on Black and White High School Students' SAT Performance. *Teachers College Record* 115 (05 2013).
[51] Pedro Domingos. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (1st. ed.). Basic Books, New York, NY.
[52] Claire Duffy. 2020. In the face of a cultural reckoning, it turns out massive corporations can move fast and fix things. *CNN Business* (June 21, 2020). https://www.cnn.com/2020/06/21/business/corporate-america-addresses-racism/index.html
[53] Claire Duffy. 2020. Plans at Microsoft and Wells Fargo to increase Black leadership are under scrutiny from the Labor Dept. *CNN Business* (October 7, 2020). https://www.cnn.com/2020/10/07/business/microsoft-wells-fargo-diverse-hiring-probe/index.html
[54] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.
[55] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics* 9, 2 (2017), 1–22.
[56] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. 2020. On Fair Selection in the Presence of Implicit Variance. In *Proceedings of the 2020 ACM Conference on Economics and Computation*.
[57] Cynthia L. Estlund. 2005. Putting Grutter to Work: Diversity, Integration, and Affirmative Action. *Berkeley J. of Labor and Employment* 26 (2005), 1.
[58] Yuri Faenza, Swati Gupta, and Xuan Zhang. 2020. Impact of Bias on School Admissions and Targeted Interventions. arXiv:2004.10846 [cs.CY]
[59] Mary J Fischer and Douglas S Massey. 2007. The effects of affirmative action in higher education. *Social Science Research* 36, 2 (2007), 531–549.
[60] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.
[61] FTC Report. January 2016. *Big data: a tool for inclusion or exclusion?* Federal Trade Commission.
[62] Rachel Goodman. 2018. Why Amazon's Automated Hiring Tool Discriminated Against Women. https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against Published Oct. 12, 2018. Last accessed Jun. 6, 2019..
[63] Alison Griswold. 2014. Why Major Companies Like Amazon Ask Job Candidates For Their SAT Scores. *Business Insider* (Mar 2014). https://www.businessinsider.com/goldman-sachs-bain-mckinsey-job-candidates-sat-scores-2014-3
[64] Craig Hanks. 2009. *Technology and values: Essential readings*. John Wiley & Sons, 7.
[65] Rema N Hanna and Leigh L Linden. 2012. Discrimination in Grading. *American Economic Journal: Economic Policy* 4 (02 2012), 146–68.
[66] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1914–1933.
[67] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.
[68] Madeline E Heilman, Caryn J Block, and Peter Stathatos. 1997. The affirmative action stigma of incompetence: Effects of performance information ambiguity. *Academy of Management Journal* 40, 3 (1997), 603–625.
[69] Kimberly A. Houser. 2019. Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making. *Stanford Tech. L. Rev.* 22 (2019), 290.

[70] Jobscan. Accessed Sept. 11, 2020. Applicant Tracking Systems. https://www.jobscan.co/applicant-tracking-systems Available at https://www.jobscan.co/applicant-tracking-systems..

[71] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. IEEE, 869–874.

[72] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[73] Michael Kearns and Aaron Roth. 2020. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (1st. ed.). oxford University Press, New York, NY.

[74] Matt Kempner. 2021. Georgia's big businesses reveal staff — and management — diversit. *The Atlanta Constitution Journal* (October 8, 2021). https://www.ajc.com/news/business/georgias-big-businesses-reveal-staff-and-management-diversity/GUZ26V3JMRCRDB2ZF4X57F3OEY/

[75] Pauline Kim. 2017. Auditing Algorithms for Discrimination. *U. Pa. L. Rev. Online* 166 (2017), 189.

[76] Pauline Kim. 2017. Data-driven discrimination at work. *Wm. & Mary L. Rev.* 58 (2017), 8657.

[77] Pauline T Kim. 2020. Manipulating opportunity. *Va. L. Rev.* 106 (2020), 867.

[78] Jon M. Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*. 33:1–33:17. https://doi.org/10.4230/LIPIcs.ITCS.2018.33

[79] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19.

[80] Mark MacCarthy. 2017. Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms. *Cumberland L. Rev.* 48 (2017), 102.

[81] James March. 1989. Exploration and Exploitation in Organizational Learning. *Organizational Science* 2 (1989), 71.

[82] Weiwen Miao and Joseph L Gastwirth. 2013. Properties of statistical tests appropriate for the analysis of data in disparate impact cases. *Law, Probability and Risk* 12, 1 (2013), 37–61.

[83] Tara Sophia Mohr. 2014. Why women don't apply for jobs unless they're 100% qualified. *Harvard Business Review* 25 (2014).

[84] Corinne Moss-Racusin, Dovidio , Victoria Brescoll, Mark Graham, and Jo Handelsman. 2012. Science Faculty's Subtle Gender Biases Favor Male Students. *Proceedings of the National Academy of Sciences* 109 (09 2012), 16474–16479. https://doi.org/10.1073/pnas.1211286109

[85] U.S. Department of Labor Office of Federal Contract Compliance Programs. 2020. U.S. DEPARTMENT OF LABOR AND MICROSOFT CORP. ENTER AGREEMENT TO RESOLVE ALLEGED HIRING DISCRIMINATION AFFECTING 1,229 APPLICANTS IN FOUR STATES. *CNN Business* (September 18, 2020). https://www.dol.gov/newsroom/releases/ofccp/ofccp20200918

[86] Richard Primus. 2010. The future of disparate impact. *Mich. L. Rev.* 108 (2010), 1341.

[87] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.

[88] Jad Salem and Swati Gupta. 2020. Under major revision at Management Science, 2021. Closing the GAP: Group-Aware Parallelization for Online Selection of Candidates with Biased Evaluations. In *International Conference on Web and Internet Economics (WINE)*. Springer.

[89] Javier Sánchez-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 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 458–468. https://doi.org/10.1145/3351095.3372849

[90] Steven D Schlachter and Jenna R Pieper. 2019. Employee referral hiring in organizations: An integrative conceptual review, model, and agenda for future research. *Journal of Applied Psychology* (2019).

[91] Oscar Schwartz. 2019. Untold History of AI: Algorithmic Bias Was Born in the 1980s. *IEEE Spectrum* (2019).

[92] Jon Shields. 2018. Over 98% of fortune 500 companies use applicant tracking systems (ATS).

[93] Gerry Smedinghoff. 2007. The art, philosophy and science of data. *Contingencies May/June* (2007), 37–40.

[94] Marion Gross Sobol and Charles J Ellard. 1988. Measures of employment discrimination: A statistical alternative to the four-fifths rule. *Industrial Relations Law Journal* (1988), 381–399.

[95] Claude M. Steele and J. Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology* 69 (1995), 797–811. Issue 5.

[96] Charles A. Sullivan. 2005. Disparate Impact: Looking Past the Desert Palace Mirage. *William & Mary Law Review* (2005).

[97] Nina Totenberg. 2020. *Supreme Court Delivers Major Victory To LGBTQ Employees.* Retrieved July 22, 2020 from https://www.npr.org/2020/06/15/863498848/supreme-court-delivers-major-victory-to-lgbtq-employees

[98] Jingyan Wang and Nihar Shah. 2019. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In *AAMAS Conference proceedings*.

[99] Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar B Shah. 2020. Debiasing Evaluations That are Biased by Evaluations. *arXiv preprint arXiv:2012.00714* (2020).

[100] Agnes Wold and Christine Wennerås. 1997. Nepotism and sexism in peer review. *Nature* 387, 6631 (1997), 341–343.

[101] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. 2020. Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 18–19.

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

A CONSTRUCTING THE POSET

As discussed in Section 5, a *poset* is a partially ranked set of candidates, and the *poset approach* is the process of forming a partial ranking and subsequently making selections based on the ranking. The purpose of this approach is to minimize the effect of bias and other inaccuracies on selection decisions. In this section, we discuss several ways in which posets can be constructed.

One natural way to construct a partial ranking of applicants is to first form score intervals (e.g., confidence intervals) based on raw scores, and then extract ordinal information from non-intersecting intervals (see, e.g., Examples 5.1 and 5.2). There are several factors that can be taken into account when forming a partial ranking in such a way. An evaluation metric itself can produce inaccurate scores, which can inform the lengths of intervals. Group-specific biases and error rates can also inform varying interval lengths by group. Evaluation metrics can also be biased against some groups, and intervals can be designed so as to mitigate known biases. To illustrate some ways in which the poset approach can be implemented, we consider an example from [88].

Example case study. In [88], Salem and Gupta analyzed the Aspiring Minds dataset, in which male and female job seekers had similar distributions of computer science test scores. As the job seekers were all in computer science fields, computer science test scores were taken as proxies for hireability. They found that by performing a linear regression on features (including gender and test scores), female[9] applicants received scores that were 16.95 points lower than those of male applicants with all other attributes equal. This does not mean that the regression model will underestimate the score of every female; rather, it means that the regression model has "learned" a trend in the data that can be harmful to some female applicants.[10]

Given that there is some data available on candidates, and a prediction mechanism that scores candidates, how can a partial ranking of the candidates be constructed in a mathematically sound way?[11] In what follows, we outline three possible approaches for constructing ranges of scores for each candidate. However, we note that other statistical approaches might be applicable as well.

1. **Data-centric approach to mitigate disparate error rates:** It is desirable that as the errors in evaluations decrease, candidates' score ranges shrink as well. This is statistically linked to the amount of data in the following way: a machine learning model trained on noisy data tends to have smaller errors as the amount of data increases. Therefore, in this approach, we will ensure that the size of the range decreases as the amount of the training data for a candidate increases. Formally, one can set the interval length to be $r(n)$, where n is the number of points in the training data and r some function that decays with n. The value of $r(n)$ should be an estimate of how much error is in the model after observing n training points, which can be calculated depending on the model class and the extent of noise; for example, when the true relationship is linear, it makes sense to choose $r(n)$ to be proportional to $1/\sqrt{n}$ [41].[12] Adopting these confidence intervals gives applicants the benefit of the doubt, giving room to account for imperfections in the evaluation mechanism.

As noted earlier, however, error rates can differ by group. The presence of an imbalanced training dataset (that is, one with an underrepresented group) can lead to higher error rates for the underrepresented group. This problem mathematically justifies having intervals of different lengths for different groups, dependent on how much uncertainty exists within each group. For example, if there were n_1 applicants from one group and n_2 applicants from another group in the training data, then one might choose intervals of length $r(n_1)$ for the former and intervals of length $r(n_2)$ for the latter.

In the context of the example case study, this would entail partitioning the pool into some number of groups and setting differing interval lengths dependent on group size. Since a bias was detected against female candidates, it may make sense to consider groups by gender, in which case the interval length for a female candidate would be proportional to $\frac{1}{\sqrt{\#\text{females}}}$, and analogously for men. There are, however, other ways to partition the pool. Each person is mapped to a point in some high-dimensional space, and clusters can be formed based on which points are close to each other with respect to some metric. We can then similarly set interval lengths based on cluster sizes.

If one insists on using equal interval lengths for all applicants, then using a length of, say, $\min\{r(n_1), r(n_2)\}$, would give each applicant the length corresponding to the error rate in the largest group. While this last method would arguably give overly narrow intervals to members of the smaller group, it still gives more benefit of the doubt to applicants (regardless of group) than using intervals of length $r(n)$.

2. **Data-centric approach to mitigate group-specific biases:** Many different approaches can fall into this category; here we highlight the *error-correction approach* used by Salem and Gupta [88]. For a given group, let y denote the vector of true scores in the training data for that group, and let $\hat{y}$ denote the predicted scores of the same applicants. Note that the transformed scores $\hat{y}_{\text{transf.}} := \frac{\sigma_y}{\sigma_{\hat{y}}}(\hat{y} - \mu_{\hat{y}}) + \mu_y$ have the same mean and standard deviation as the true scores y. The errors of each of these transformed scores are thus in the vector $y - \hat{y}_{\text{transf.}}$, the standard deviation σ of which measures the overall error of $\hat{y}_{\text{transf.}}$. One can there use score ranges of

$$\left[\hat{y}_{\text{transf.}}(e) - \lambda\sigma,\ \hat{y}_{\text{transf.}}(e) + \lambda\sigma\right]$$

for each applicant e in the given group, where λ is a parameter representing the desired level of confidence in the score range.

3. **Human-centric approach:** When there is direct human involvement in evaluating applicants (e.g., through interviews), biases

[9] All entries in the dataset listed a gender of male or female.

[10] One of the reasons for this might be imbalance of data points in the male versus female group. For example, the average score in the training set among all female applicants was 455.30, and the average score among all male applicants was 478.05 (a slightly larger discrepancy than in the entire dataset).

[11] In the example above, the scoring mechanism was a simple linear regression. But in general, it can involve natural language processing of the text in résumés, a neural network built on top of that, or any other machine learning technique.

[12] If the true relationship between a candidate's features and hireability is linear, and the noise is independently drawn from identical normal distributions, then the linear regression estimation of hireability is unbiased and has standard deviation proportional to $1/\sqrt{n}$. The proportionality constant depends on the distribution of the datapoints and the variance of the noise [41].

and inaccuracies might still persist. One might think that the poset approach would not help in this situation, but it can. Similar to the data-centric approach, one can reduce this error through repeated independent evaluations (i.e., wisdom of crowds [93]). In this case, a single applicant can be interviewed by a diverse committee, each of whose members scores the applicant. These scores can then be used to form score ranges for each applicant (e.g., the minimum score to the maximum score, or the first to the third quartile, etc.). An applicant's score range can be decreased through further discussion by the committee until the interval is sufficiently small to allow for reasonably many comparisons in the applicant pool.[13] If there were some way to obtain repeated evaluations (e.g., through multiple human evaluators, or access to multiple scoring algorithms) of the job seekers in the example case study, then this human-centric approach could be applied there as well. For each person in the dataset, we could repeatedly obtain evaluations until we can be confident in their score (say, up to 5 points). If certain groups (e.g., gender groups, or clusters obtained by a machine learning algorithm) experience higher variance in evaluations, then the same number of evaluations might result in confidence intervals of different lengths for different groups.

Accounting for other sources of errors. There are sources of errors which are intrinsic to machine learning and predictions, and interval lengths can also be designed to mitigate these errors. For example, if a highly informative or causal variable is absent from the data, then predictions can suffer (this type of error is called *Bayes error*). Another type of error, called *approximation error*, describes error resulting from the mismatch between the true relationship between applicant features and ability, and the class of models that can be produced by the algorithm. Both of these types of errors are independent of the size of the training data, so we generally do not expect these errors to decay with time. In an attempt to account for this sort of intrinsic error, one could impose a minimum interval length for all candidates, where the minimum length depends on the accuracy of the evaluation metric.

B POSET APPROACH DIAGRAMS

In Section 5, we discussed how the poset approach can be used in the screening of applications. To determine the appropriateness of this approach to screening, it is of legal, ethical, and utilitarian importance to understand the effect of the poset approach on applicants. While the practical effect of the poset approach will depend on context and precise implementation, it is informative to observe its effect on artificial data.

The four examples shown in this section (Figures 4-7) compare candidate slates produced by a cutoff on raw scores versus on score ranges. To illustrate the effect of using score ranges over groups, we consider two groups, where Group 2 has a lower mean evaluation and a larger interval length (perhaps due to dataset imbalance issues as discussed earlier). In Figure 4, we see how using score ranges instead of raw scores can increase the selection rate of the minority group. However, the use of score ranges does not necessarily benefit the minority group in general. Figures 5-7 consider the same scenario as Figure 4 but with different distributional assumptions. Of note is how the use of score ranges can benefit the majority or minority group, and can benefit the low-scoring or the high-scoring group. This point, in particular, means that the poset approach does not inherently constitute a quota system, which is an important feature with respect to anti-discrimination law (see Section 6.2).

[13] Such practices are already prevalent within hiring committees and program committees for conferences such as ICLR, NeurIPS and WWW.

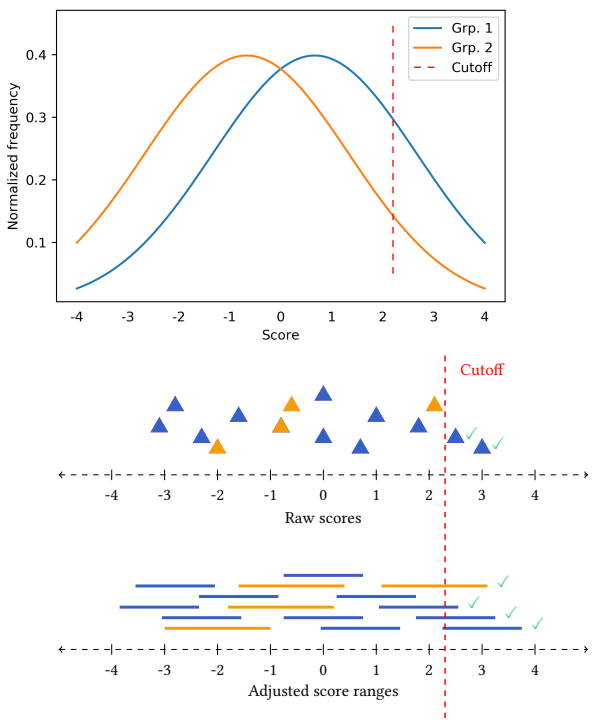

Figure 4: (top) Example of score distributions (blue: Group 1, orange: Group 2) and (bottom) potential score ranges for candidates from these distributions. Suppose a hiring committee wants to select two of the applicants represented in the bottom plot. If only the raw evaluations (the centers of the intervals) are used to make these decisions, then only the two high-scoring Group 1 candidates could be selected, as they are the only applicants meeting the cutoff. However, if score ranges are considered, then the highest-scoring Group 2 candidate meets the cutoff as well. In this example, adopting the poset method results in a more diverse slate of candidates meeting the cutoff, vis-à-vis using raw scores.

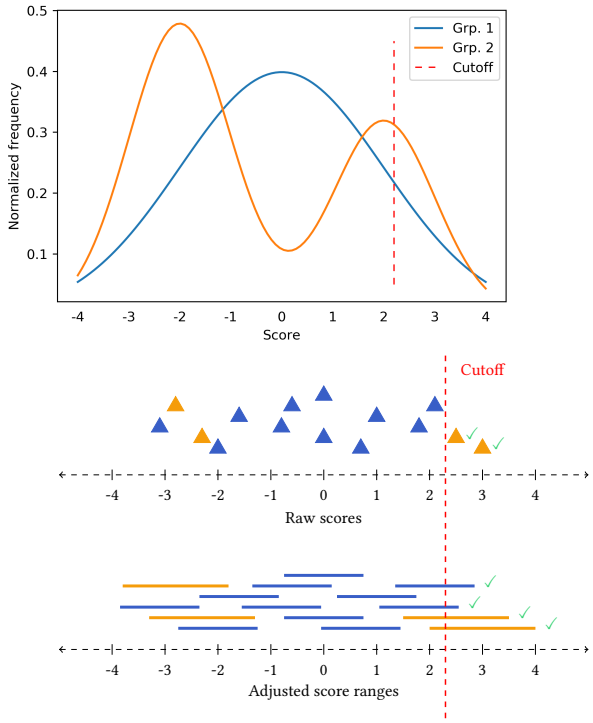

Figure 5: (top) Example of (somewhat unusual) score distributions (blue: Group 1, orange: Group 2) and (bottom) potential score ranges for candidates from these distributions. Suppose a hiring committee wants to select two of the applicants corresponding to the bottom plot. If only the raw evaluations (the centers of the intervals) are used to make these decisions, then only the two high-scoring Group 2 candidates could be selected, as they are the only applicants meeting the cutoff. However, if score ranges are considered, then the two highest-scoring Group 1 candidates meet the cutoff as well. From this example, we see that adopting the poset approach can be beneficial to the majority group as well and does not routinely advantage the lower-mean group (in this case, Group 2).

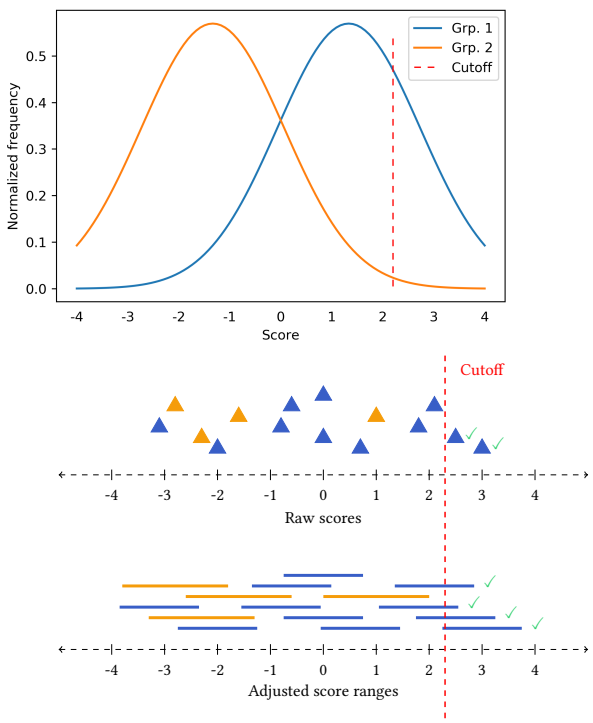

Figure 6: (top) Example of score distributions (blue: Group 1, orange: Group 2) and (bottom) potential score ranges for candidates from these distributions. Suppose a hiring committee wants to select two of the applicants corresponding to the bottom plot. If only the raw evaluations (the centers of the intervals) are used to make these decisions, then only the two highest-scoring Group 1 candidates could be selected, as they are the only applicants meeting the cutoff. If the score ranges are considered, then the four highest-scoring Group 1 candidates meet the cutoff. This example shows that adopting the poset approach does not necessarily increase the selection rate for the group with the lower mean score, and that the poset approach does not necessarily constitute a quota system.

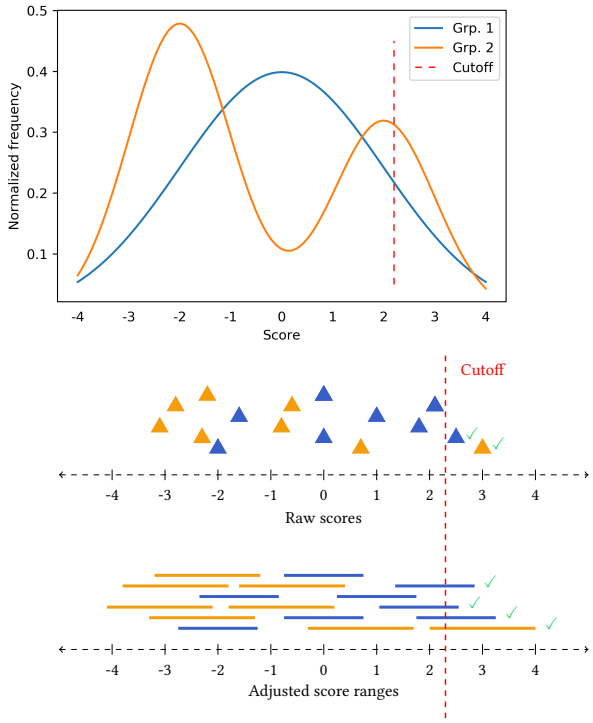

Figure 7: (top) Example of (somewhat unusual) score distributions (blue: Group 1, orange: Group 2) and (bottom) potential score ranges for candidates from these distributions. Suppose a hiring committee wants to select two of the applicants corresponding to the bottom plot. If only the raw evaluations (the centers of the intervals) are used to make these decisions, then one Group 1 and one Group 2 candidate will be selected, as they are the only applicants meeting the cutoff. In this case, demographic parity is achieved, as both groups have equal size. However, if score ranges are considered, then two additional Group 1 candidates meet the cutoff as well. This example shows that adopting the poset approach does not necessarily make the new candidate slate (i.e., those meeting the cutoff) more representative compared to using raw scores—indeed, in this example, adopting the poset approach moves the new candidate slate farther away from demographic parity.