Interventions for Ranking in the Presence of Implicit Bias

L. Elisa Celis
Yale University

Anay Mehrotra
IIT Kanpur

Nisheeth K. Vishnoi
Yale University

ABSTRACT

Implicit bias is the unconscious attribution of particular qualities (or lack thereof) to a member from a particular social group (e.g., defined by gender or race). Studies on implicit bias have shown that these unconscious stereotypes can have adverse outcomes in various social contexts, such as job screening, teaching, or policing. Recently, [33] considered a mathematical model for implicit bias and showed the effectiveness of the Rooney Rule as a constraint to improve the utility of the outcome for certain cases of the subset selection problem. Here we study the problem of designing interventions for the generalization of subset selection – ranking – that requires to output an ordered set and is a central primitive in various social and computational contexts. We present a family of simple and interpretable constraints and show that they can optimally mitigate implicit bias for a generalization of the model studied by Kleinberg and Raghavan. Subsequently, we prove that under natural distributional assumptions on the utilities of items, simple, Rooney Rule-like, constraints can also surprisingly recover almost all the utility lost due to implicit biases. Finally, we augment our theoretical results with empirical findings on real-world distributions from the IIT-JEE (2009) dataset and the Semantic Scholar Research corpus.

CCS CONCEPTS

• **Information systems** → *Content ranking*; • **Mathematics of computing** → *Probability and statistics*;

KEYWORDS

Implicit bias, ranking, algorithmic fairness, interventions

ACM Reference Format:
L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2020. Interventions for Ranking in the Presence of Implicit Bias. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3351095.3372858

1 INTRODUCTION

Implicit bias is the unconscious attribution of particular qualities (or lack thereof) to a member from a particular social group defined by characteristics such as gender, origin, or race [26]. It is well understood that implicit bias is a factor in adverse effects against subpopulations in many societal contexts [1, 6, 41] as also highlighted by recent events in the popular press [21, 37, 60]. For instance, in employment decisions, men are perceived as more competent and given a higher starting salary even when qualifications are the same [51], and in managerial jobs, it was observed that women had to show roughly twice as much evidence of competence as men to be seen as equally competent [36, 58]. In education, implicit biases have been shown to exist in ways that exacerbate the achievement gap for racial and ethnic minorities [52] and female students [40], and add to the large racial disparities in school discipline which particularly affect black students' school performance and future prospects [44]. Beyond negatively impacting social opportunities, implicit biases have been shown to put lives at stake as they are a factor in police decisions to shoot, negatively impacting people who are black [19] and of other racial or ethnic minorities [47]. Furthermore, decision making that relies on biased measures of quantities such as utility can not only adversely impact those perceived more negatively, but can also lead to sub-optimal outcomes for those harboring these unconscious biases.

To combat this, a significant effort has been placed in developing anti-bias training with the goal of eliminating or reducing implicit biases [23, 38, 63]. However, such programs have been shown to have limited efficacy [43]. Furthermore, as algorithms increasingly take over prediction and ranking tasks, e.g., for ranking and shortlisting candidates for interview [8], algorithms can learn from and encode existing biases present in past hiring data, e.g., against gender [20] or race [53], resulting in algorithmic biases of their own. Hence, it is important to develop interventions that can mitigate implicit biases and hence result in better outcomes.

As a running example we will consider hiring, although the interventions we describe would apply to any domain in which people are selected or ranked. Hiring usually works in multiple stages, where the candidates are first identified and ranked in order of their perceived relevance, a shortlist of these candidates are then interviewed with more rigor, and finally a one or few of them are hired [8]. The Rooney Rule is an intervention to combat biases during the shortlisting phase, it requires the "shortlister" (an entity that shortlists) to select at least one candidate from an underprivileged group for interview. It was originally introduced for coach positions in the National Football League [18], and subsequently adopted by other industries [11, 27, 45, 49]. The idea is that including the underprivileged candidates would give opportunity to these candidates, with a higher (hidden or latent) potential. Whereas without the Rooney Rule these candidates may not have been selected for interview. While the Rooney Rule appears to have been effective[1] it is just one of many possible interventions one could design. How can we theoretically compare two proposed interventions?

[33] study the Rooney Rule under a theoretical model of implicit bias, with two disjoint groups $G_a, G_b \subseteq [m]$ of m candidates, where

[1] The representation of African-American coaches in NFL increased from 6% to 22% since Rooney Rule's introduction in 2002 to 2006 [18].

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

ACM ISBN 978-1-4503-6936-7/20/02...$15.00
https://doi.org/10.1145/3351095.3372858

G_b is the underprivileged group. Each candidate $i \in [m]$ has a true, *latent* utility $w_i \in \mathbb{R}$, which is the utility they would generate if hired, and an *observed* utility $\hat{w}_i \leq w_i$, which is the shortlister's (possibly biased) estimate of w_i. The shortlister selects $n \leq m$ candidates with the highest observed utility. For example, in the context of peer-review, the latent utility of a candidate could be their total publications, and the observed utility could be the total weight the reviewer assigns to the publications ("impact points" in [56]). They model implicit bias as a multiplicative factor $\beta \in [0, 1]$, where the observed utility is $\hat{w}_i \coloneqq \beta \cdot w_i$ if $i \in G_b$, and $\hat{w}_i \coloneqq w_i$ if $i \in G_a$. [33] characterize conditions on n, m, β and the distribution of w_i, where Rooney rule increases the total utility of the selection.

However, before the shortlisting phase, applicants or potential candidates must be identified and ranked. For example, LinkedIn Recruiter predicts a candidate's "likelihood of being hired" from their activity, Koru Hire analyzes a candidates (derived) personality traits to generate a "fit score", and HireVue "grades" candidates to produce an "insight score" [8]. In a ranking, the candidates are ordered (in lieu of an intervention, by their observed utilities), and the utility of a ranking is defined by a weighted sum of the latent utilities of the ranked candidates where the weight decreases the farther down the ranking a candidate is placed. Such weighting when evaluating rankings is common practice, and due to the fact that candidates placed lower in the list receive a lower attention as compared to those placed higher [30]; this translates into being less likely to be shortlisted, and contributing less to the total utility of the hiring process. Therefore, it becomes important to consider interventions to mitigate bias in the ranking phase, and understand their effectiveness in improving the ranking's latent utility.

Can we construct simple and interpretable interventions that increase the latent utility of rankings in the presence of implicit bias?

1.1 Our contributions

We consider the setting where items to be ranked may belong to multiple intersectional groups and present a straightforward generalization of the implicit bias model in [33] for this setting. We consider a set of interventions (phrased in terms of constraints) for the ranking problem which require that a fixed number of candidates from the under privileged group(s) be represented in the top k positions in the ranking for all k. We show that for any input to the ranking problem – i.e., any set of utilities or bias – there is an intervention as above that leads to optimal latent utility (Theorem 3.1). We then prove a structural result about the optimal ranking when all the utilities are drawn from the same distribution, making no assumption on the distribution itself (Theorem 3.2). This theorem gives a simple Rooney Rule-like intervention for ranking in the case when there are two groups: For some α, require that in the top k positions, there are at least $\alpha \cdot k$ items from the underprivileged group. We then show that when the utilities are drawn from the uniform distribution, the latent utility of a ranking that maximizes the biased utility but is subject to our constraints (for an appropriate α) is close to that of the optimal latent utility (Theorem 3.3). We evaluate the performance of ranking under our constraints empirically on two real-world datasets, the IIT-JEE 2009 dataset (see Section 4) and the Semantics Scholar dataset (see Section E.2 in the Supplementary Material). In both cases we observe that our simple constraints significantly improve the latent utility of the ranking, and in fact attain near-optimal latent utility. Finally, while we phrase these results in the context of implicit bias, such interventions would be effective whenever the observed utilities are systematically biased against a particular group.

1.2 Related work

There is a large body of work on studying the effects of bias in rankings, and designing algorithms for 'fair rankings'; see, e.g., [17, 34, 42, 48, 50]. We refer the reader to the excellent talk by Castillo [10] that gives an overview of work on fairness and transparency in rankings. A significant portion of these works are devoted to generating unbiased rankings. For example, several approaches strive to learn the latent utilities of items and output a ranking according to the learned values [4, 61]. In contrast, we do not strive to learn the latent utilities, rather, to find a ranking that is close to that given by the (unknown) ranking according to the (unknown) latent utilities. A different approach instead considers constraints on the placement of individuals within the ranking depending on their group status, e.g., enforcing that at least x% of the top k candidates be non-male [17, 24]. These works take the constraints as input and develop algorithms to find the optimal feasible ranking. While we also use constraints in our approach, our goal differs; we strive to show that such constraints can recover the optimal latent utility ranking, and, where possible, derive the appropriate constraints that achieve this.

Constraints which guarantee representation across various groups have been studied in a number of works on fairness across various applications [14–16, 28], most relevantly in works on forms of subset selection [13] and ranking [17, 24, 62]. The primary goals of these works is to design fast algorithms that satisfy the constraints towards satisfying certain definitions of fairness; these fairness goals are given exogenously and the utilities are assumed to be unbiased. In contrast, we begin with the premise that the utilities are systematically incorrect due to implicit bias, and use the constraints to mitigate the effect of these biases when constructing a ranking. Our goal is to determine how to construct effective interventions rather than on the algorithm for solving the constraints; in fact, we use some of the works above as a black box for the subroutine of finding a ranking once the constraints have been determined.

Studying implicit and explicit biases is a rich field in psychology [36, 47, 58], where studies propose several mechanisms for origins of biases [46] and analyze present-day factors which can influence them [22]. We point the reader to Greenwald and Banaji's seminal work on implicit biases [25], and the excellent treatise by Whitley and Kite [57] for an overview of the field. We consider one model of implicit bias inspired by [33, 56]; however other relevant models may exist and exploring other kinds of bias and models for capturing them could lead to interesting expansions of this work.

2 RANKING PROBLEM, BIAS MODEL, AND INTERVENTIONS

2.1 Ranking problem

In the classical ranking problem, one is given m items, where item i has utility w_i, from which a ranked list of $n \leq m$ items has to be outputted. A ranking is a one to one mapping from the set of items $[m]$ to the set of positions $[n]$. It is sometimes convenient to let $x \in \{0,1\}^{m\times n}$ denote a binary assignment matrix representing a ranking, where $x_{ij} = 1$ if the i-th item is placed at position j, and is 0 otherwise. Define a position-based discount $v \in \mathbb{R}^n_{\geq 0}$, where an item placed at position $j \in [n]$, contributes a latent utility of $w_i \cdot v_j$. The latent utility obtained by placing an item i at position j is then $w_i v_j$. It is assumed that $v_j \geq v_{j+1}$ for all $1 \leq j \leq n-1$ implying that the same item derives a higher utility at a lower position. This is satisfied by popularly studied position-based discounts such as discounted cumulative gain (DCG) [30] where $v_k := 1/\log(k+1)$ (and its variants) and Zipfian where $v_k := 1/k$ [32]. Then, given $v \in \mathbb{R}^n_{\geq}$ we define the *latent* utility of a ranking x as

$$\mathcal{W}(x,v,w) := \sum_{i\in[m],\ j\in[n]} x_{ij} w_i v_j. \quad \text{(Latent utility, 1)}$$

The goal of the ranking problem is to find a ranking (equivalently, an assignment) that maximizes the latent utility:

$$\operatorname{argmax}_x \mathcal{W}(x,v,w).$$

The reason the utility above is called "latent" is, as is shortly discussed, in the presence of implicit bias, the perceived utility may be different from the latent utility. Note that subset selection is a special case of the ranking problem when $v_j = 1$ for all $j \in [n]$.

2.2 Groups and a model for implicit bias

Items may belong to one or more of intersectional (i.e., not necessarily disjoint) groups $G_1, G_2, \ldots, G_p$. Each $G_s \subseteq [m]$ and $G_s \cap G_t$ may not be empty for $s \neq t$. The perceived or observed utility of group items in G_s may be different from their latent utility. And, items that belong to multiple groups may be subject to multiple implicit biases, as has been observed [7, 51]. To mathematically model this, we consider a model of implicit bias introduced by by [33], which is motivated from empirical findings of [56]: For two disjoint groups, $G_a, G_b \subseteq [m]$, given an implicit bias parameter $\beta \in [0,1]$ they defined the observed utility as

$$\hat{w}_i := \begin{cases} w_i & \text{if } i \in G_a \\ \beta w_i & \text{if } i \in G_b. \end{cases} \quad \text{(Observed utility, 2)}$$

To extend this model to the case of multiple intersectional groups, for each $s \in [p]$, we assume an implicit bias parameter $\beta_s \in [0,1]$. Since, it is natural to expect that items at the intersection of multiple groups encounter a higher bias [58], we define their implicit bias parameter as the product of the implicit biases of each group the item belongs to. Formally, the observed utility $\hat{w}_i$ of item $i \in [m]$ is

$$\hat{w}_i := \left(\prod_{s\in[p]:\ G_s \ni i} \beta_s \right) \cdot w_i. \quad (3)$$

It follows that the case of two disjoint groups G_a, G_b is a special case; let $\beta_a = 1$ and $\beta_b = \beta$.

2.3 Intervention constraints

In the presence of implicit bias, the ranking problem then results in finding the assignment matrix x that maximizes the observed utility $\mathcal{W}(x,v,\hat{w})$. Thus, not only does it result in adverse outcomes for groups for which $\beta_s < 1$, it also follows that optimizing this utility as such may be sub-optimal for the overall goal of finding a utility maximizing rank. To see this note that if $x^\star$ is the assignment that maximizes $\mathcal{W}(x,v,\hat{w})$, the value of the latent utility, $\mathcal{W}(x^\star,v,w)$, derived from it may be much less.

Motivated by the Rooney Rule and its efficacy as an intervention in the subset selection problem [33], we investigate if there are constraints that can be added to the optimization problem of finding a ranking that maximizes the observed utilities, that results in a ranking in which the latent utility is much higher, possibly even optimal: $\max_x \mathcal{W}(x,v,w)$. As a class of potential interventions, we consider lower bound constraints on the number of items from a particular group $s \in [p]$, selected in the top-k positions of the ranking, for all positions $k \in [n]$. More specifically, given $L \in \mathbb{Z}^{n\times p}_{\geq 0}$ we consider the following constraints on rankings (assignments):

$$\forall\, k \in [n],\ s \in [p] \quad L_{ks} \leq \sum_{j\in[k]} \sum_{i\in G_s} x_{ij}. \quad (4)$$

We will sometimes refer to these constraints as L-constraints. Let

$$\mathcal{K}(L) := \{x \in \{0,1\}^{m\times n} \mid x \text{ satisfies } L\text{-constraints}\} \quad (5)$$

be the set of all rankings satisfying the L-constraint. Our goal will be to consider various L-constraints and understand under what conditions on the input utilities and bias parameters does the ranking

$$\tilde{x} := \operatorname{argmax}_{x\in\mathcal{K}(L)} \mathcal{W}(x,v,\hat{w}) \quad (6)$$

have the property that $\mathcal{W}(\tilde{x},v,w)$ close to $\max_x \mathcal{W}(x,v,w)$.

Constraints such as (4) have been studied by a number of works on fairness, including by several works on ranking [17, 24, 62]. While these constraints can encapsulate a variety of fairness and diversity metrics [12], their effectiveness as an intervention for implicit biases was not clear and the utility of the rankings generated remained ill-understood prior to this work.

3 THEORETICAL RESULTS

Notation. Let Z be a random variable. We use $Z_{(k:n)}$ to represent the k-th order statistic (the k-th largest value) from n iid draws of Z. For all $a < b$, define $\mathcal{U}[a,b]$ to be the uniform distribution on $[a,b]$. More generally, for an interval $I \subseteq \mathbb{R}$, let $\mathcal{U}I$ be the uniform distribution on I. Let

$$x^\star := \operatorname{argmax}_x \mathcal{W}(x,v,w) \quad (7)$$

be the ranking that maximizes the latent utility.

3.1 L-constraints are sufficient to recover optimal latent utility

Our first result is structural and shows that the class of L-constraints defined above are expressive enough to recover the optimal latent utility while optimizing observed utility constrained to certain specific $L \in \mathbb{Z}^{n\times p}_{\geq 0}$.

Theorem 3.1. *Given a set of latent utilities $\{w_i\}_{i=1}^m$, there exists constraints $L(w) \in \mathbb{Z}^{n\times p}_{\geq 0}$, such that, for all implicit bias parameters $\{\beta_s\}_{s=1}^p \in (0,1)^p$, the optimal constrained ranking $\tilde{x} :=$ $\operatorname{argmax}_{x\in\mathcal{K}(L(w))} \mathcal{W}(x,v,\hat{w})$ satisfies*

$$\mathcal{W}(\tilde{x},v,w) = \max_x \mathcal{W}(x,v,w). \quad (8)$$

Without additional assumptions, $L(w)$ necessarily depends on w; see Fact D.1 in the Supplementary material. A set of utility-independent constraints is often preferable due to its simplicity and interpretability; our next two results take steps in this direction by making assumptions about distributions from which the utilities are drawn.

Proof sketch of Theorem 3.1. Consider the following constraints: For all $s \in [p]$ and $k \in [n]$,

$$L_{ks}(w) := \sum_{i \in G_s, j \in [k]} x^{\star}_{ij}.$$

Recall that $x^{\star} := \operatorname{argmax}_x \mathcal{W}(x, v, w)$ is a function of w. We claim that $L_{ks}(w)$ satisfy the claim in the theorem. The proof proceeds in two steps. First, we show that $\tilde{x}$ is the same as $x^{\star}$ up to the groups of items at each position. Let $T_i := \{s \colon i \in G_s\}$ be the set of groups i belongs to. This proof relies on the fact that the observed utility of an item is always smaller than its latent utility, and that for any two items $i_1, i_2 \in [m]$, if $T_{i_2} \subsetneq T_{i_1}$ and $w_{i_1} = w_{i_2}$, then $\hat{w}_{i_1} < \hat{w}_{i_2}$. Using these we show that for each position $k \in [n]$, under the chosen L-constraints, it is optimal to greedily place item $i' \in [m]$ that has the highest observed utility and satisfies the constraints. In the next step, we show that $\tilde{x}$ has the same latent utility as $x^{\star}$ from a contradiction. We show that if the claim is satisfied for the first $k \in [n]$ positions, then we can swap two candidates i_1 and i_2, such that $T_{i_1} = T_{i_2}$, to satisfy the claim for the first $(k+1)$ positions without loosing latent utility. Here we use the fact for any two items $i_1, i_2 \in [m]$, if $T_{i_2} = T_{i_1}$ and $w_{i_1} < w_{i_2}$, then $\hat{w}_{i_1} < \hat{w}_{i_2}$, i.e., the relative order of items in the same set of groups does not change whether we rank them by their observed utility (as in $\tilde{x}$) or their latent utility (as in $x^{\star}$).[2] The proof of Theorem 3.1 is presented in Section 7.1.

3.2 Distribution independent constraints

We now study the problem of coming up with constraints that do not depend on the utilities. Towards this, we consider the setting of two disjoint groups, $G_a, G_b \subseteq [m]$. Let $m_a := |G_a|$ and $m_b := |G_b|$ be the sizes of G_a and G_b. We assume that the latent utility W_i for all items $i \in G_a \cup G_b$ is i.i.d. and drawn from some distribution $\mathcal{D}$. This model is equivalent to the one considered by [33], except that they fix $\mathcal{D}$ to be in the family of power-law distributions, whereas our result in this section holds for any distribution $\mathcal{D}$.

The optimal ranking will sort the utilities $(w_i)_{i \in [m]}$ in a decreasing order (breaking ties arbitrarily if necessary). For all $\ell \in [m_b]$, let $P_\ell \in [m]$ be the random variable representing the position of the ℓ-th item from G_b in the optimal ranking. Let N_{kb} be the random variable that counts the number of items belonging to G_b in the first k positions of the optimal ranking. The following result reveals the structure of the optimal ranking (when there is no implicit bias) and is used in the next subsection to design utility-independent constraints.

Theorem 3.2. *Let $\mathcal{D}$ be a continuous distribution, $\ell \leq m_b$, and $0 < k < \min(m_a, m_b)$ be a position, then*

$$\forall\, \delta \geq 2 \qquad \Pr[\, N_{kb} \leq \mathbb{E}[N_{kb}] - \delta \,] \leq e^{-\frac{2(\delta^2-1)}{k}}, \tag{9}$$

$$\mathbb{E}[N_{kb}] = k \cdot {}^{m_b}/_{(m_a+m_b)}, \tag{10}$$

$$\mathbb{E}[P_\ell] = \ell \cdot \left(1 + {}^{m_a}/_{(m_b+1)}\right). \tag{11}$$

Note that this result is independent of the distribution $\mathcal{D}$ and only requires $\mathcal{D}$ to be a continuous probability distribution. The above equations show that with high probability the optimal ranking has $k \cdot {}^{m_b}/_{(m_a+m_b)}$ items from G_b in the top-k positions, and in expectation, it places these items at equal distances in the first k positions for all $k \in [n]$. This observation motivates the following simple constraints.

Simple constraints. Given a number $\alpha \in [0, 1]$, we define the constraints $L(\alpha)$ as follows: For all $k \in [n]$

$$L_{ka} := 0 \text{ and } L_{kb} := \alpha k. \qquad \text{(Rooney Rule like constraints, 12)}$$

Note that the only non-trivial constraint is on G_b. These constraints are easy to interpret and can be seen as generalization of the Rooney Rule to the ranking setting.

Proof sketch of Theorem 3.2. Here, we discuss the distributional independence of Theorem 3.2, and present its proof in Section 7.2. For all $i \in [m]$, $w_i \overset{d}{=} \mathcal{D}$ be the random utility of the i-th item drawn from $\mathcal{D}$ and $F_{\mathcal{D}}(\cdot)$ be the cumulative distribution function of $\mathcal{D}$. Then the independence follows from the straightforward facts that for all $\mathcal{D}$, $F_{\mathcal{D}}(w_i) \overset{d}{\sim} \mathcal{U}[0, 1]$ (here is one place it is used that $\mathcal{D}$ is continuous) and that $F_{\mathcal{D}}(\cdot)$ (being a cdf) is a monotone function. From these it follows that $\operatorname{argmax}_x \mathcal{W}(x, v, \{w_i\}_{i=1}^m) = \operatorname{argmax}_x \mathcal{W}(x, v, \{F_{\mathcal{D}}(w_i)\}_{i=1}^m)$. Thus, we can replace the all W_is by $F_{\mathcal{D}}(w_i)$ without changing the optimal ranking.

3.3 Optimal latent utility from simple constraints for uniform distributions

In this section we study the effect of the simple constraints mentioned in the previous subsection when $\mathcal{D}$ is the uniform distribution on $[0, 1]$, for the setting of two disjoint groups $G_a, G_b \subseteq [m]$ with $\beta_a = 1$ and $\beta_b = \beta$.

We discuss how these arguments could be extended to other bounded distributions in the remarks following Theorem 3.3, and empirically study the increase in utility for a non-bounded distribution in Section 4.

Define the expected utility $U_{\mathcal{D},v}(\alpha, \beta)$ as

$$U_{\mathcal{D},v}(\alpha, \beta) := \mathop{\mathbb{E}}_{w \leftarrow \mathcal{D}^m} \left[\mathcal{W}(\tilde{x}, v, w)\right] \tag{13}$$

where for each draw w, $\tilde{x} := \operatorname{argmax}_{x \in \mathcal{K}(L(\alpha))} \mathcal{W}(x, v, \hat{w})$, and $\hat{w}_i = w_i$ if $i \in G_a$ and $\hat{w}_i = \beta w_i$ if $i \in G_b$. Sometimes we drop the subscripts $\mathcal{D}$ and v if they are clear from the context.

Theorem 3.3. *Given a $\beta \in (0, 1)$, if $\mathcal{D} := \mathcal{U}[0, 1]$, v satisfies Assumptions (14) and (15) with $\varepsilon > 0$, and $n \leq \min(m_a, m_b)$, then for $\alpha^{\star} := \frac{m_b}{m_a + m_b}$, then adding $L(\alpha)$-constraints achieve nearly optimal latent utility in expectation:*

$$U_{\mathcal{D},v}(\alpha^{\star}, \beta) = U_{\mathcal{D},v}(0, 1) \cdot (1 - O(n^{-\varepsilon/2} + n^{-1})).$$

$$\frac{\sum_{k=1}^{n-1} v_k - v_{k+1}}{\sum_{k=1}^{n} v_k} = \frac{v_1 - v_n}{\sum_{k=1}^{n} v_k} = O(n^{-\varepsilon}). \tag{14}$$

$$\forall\, k \in [n],\; v_k - v_{k+1} \geq v_{k+1} - v_{k+2}. \tag{15}$$

Roughly, these assumptions mean that the position discounts be much larger than the difference between between the discounts of two consecutive positions. These are mild assumptions, and are satisfied by several commonly studied position discounts including

[2] A minor technicality is that $\tilde{x}$ could swap two items i_1, $i_2 \in [m]$, with $T_{i_2} = T_{i_1}$ and $w_{i_1} = w_{i_2}$, relative to $x^{\star}$. But this does neither affects the latent utility nor the observed utility.

DCG for $\varepsilon = 0.9\bar{9}$, and inverse polynomial discount where $v_k := k^{-c}$ and $c \in (0,1)$ [55] for $\varepsilon = 1 - c$.

A few remarks are in order:

(1) When $v_k = 1$ for all $k \in [n]$ then we can derive the following explicit expressions of the utility assuming $m_a, m_b \geq n$:

$$U_{\mathcal{D},1}(\alpha^\star, \beta) \qquad \text{(Utility with constraints, 16)}$$
$$= n\Big(1 - \frac{n}{2(m_a + m_b)} + O\big(n^{-3/8}\big)\Big),$$
$$U_{\mathcal{D},1}(0, \beta) \qquad \text{(Utility without constraints, 17)}$$
$$= \begin{cases} \frac{m_a(1-\beta^2)}{2} + \frac{m_a\beta^2+m_b}{2}\Big[1 - \frac{(m_a+m_b-n)^2}{(m_a\beta+m_b)^2} + O\big(n^{-3/8}\big)\Big], & c = n - \omega(n^{5/8}) \\ n\big(1 - \frac{n}{2m_a} + O\big(n^{-3/8}\big)\big), & c \geq n + \Theta(n^{5/8}). \end{cases}$$

Where, we define $c := m_a(1-\beta)$.

(2) Choosing $\beta = 1$ in Equation (17) we can see that the optimal latent utility from a ranking of n items is $n(1 - n/(2m_a+2m_b) + o(1))$, and that the constraints achieve this utility within a $(1 - \Theta(n^{-1}))$ multiplicative factor *for any* $\beta \in (0,1)$. Therefore, for all $0 < \beta < 1$ there exists an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, the constraints achieve nearly optimal latent utility.

Remark 3.4. *We note that our choice of* $\alpha^\star := \frac{m_b}{m_a+m_b}$ *is independent of* β*. Thus, we can achieve near optimal latent utility without the knowledge of* β*.*

(3) Extending $\mathcal{D}$ to $\mathcal{U}[0, C]$, does not change the form of the theorem. We can simply scale Equations (16) and (17) by C.

(4) In the special case of subset selection ($v_k = 1$ for all k), this theorem answers the open problem in [33] regarding the ℓ-th order Rooney rule for the uniform distribution.

Proof sketch of Theorem 3.3. To calculate $U_{\mathcal{D},v}(0,\beta)$, we partition the items from G_a into those with a "high" utility (in $(\beta, 1]$) and all others (with utility in $[0, \beta]$). Items with a high utility are always selected before any item from G_b. The number of such items, N_{a_1}, is a sum n random variables indicating if the utility is in $[\beta, 1)$. Therefore, N_{a_1} has a binomial distribution Binomial$(n, 1-\beta)$. Conditioned on N_{a_1}, we can show that the distribution of observed utilities of all remaining items, including those from G_b, is the same. This uses the fact that a uniform random variable conditioned to lie in an sub-interval is uniform. Using this symmetry we can show that the number of items selected from G_b, N_b, follows a hypergeometric distribution conditioned on N_{a_1}.

This gives us a value for $U_{\mathcal{D},v}(0,\beta)$ conditioned on N_{a_1}. To do away with the conditioning, we derive approximations to the negative moments of the binomial distribution.

To upperbound the difference $U_{\mathcal{D},v}(0,1) - U_{\mathcal{D},v}(\alpha^\star,\beta)$, we use a coupling argument and the concentration properties of the hypergeometric distribution to show that the difference in positions of any item i between the constrained and the unconstrained ranking is $o(n^\delta)$ with high probability. Using Assumptions (14) and (15) with the boundedness of the utility gives us a lower bound of

$$U_{\mathcal{D},v}(0,1) - U_{\mathcal{D},v}(\alpha^\star,\beta) = O(n^{\delta-\varepsilon} + n^{-1}) \sum\nolimits_{k=1}^{n} v_k.$$

Further, using the fact that $U_{\mathcal{D},v}(0,1) = \Omega(\sum_{k=1}^{n} v_k)$, the theorem follows by choosing $\delta = \varepsilon/2$.

To find an explicit expression of $U_{\mathcal{D},v}(\alpha,\beta)$ in the special case when $v_k = 1$, we use a coupling argument to show that if the unconstrained ranking picks fewer than $n/2$ items from G_b then the constrained selects exactly $n/2$ items from both G_a and G_b. Using the distribution of N_b for the unconstrained case we can show that this occurs with high probability as long as $\beta < 1$ (note the strict inequality). Then, using the boundedness of the utility, we show that $U_{\mathcal{D},v}(\alpha^\star,\beta)$ is twice the sum of expected utility of the highest $n/2$ order statistics in n draws from $\mathcal{U}[0,1]$ which gives us the required expression. The proof of Theorem 3.3 in presented in Section B of the supplementary material.

Remark 3.5. *We expect similar strategy would give us bounds on* $U_{\mathcal{D},v}(0,\beta)$ *and* $U_{\mathcal{D},v}(\alpha^\star,\beta)$ *with other bounded distributions as well. We provide some evidence in favor of this for a naturally occurring bounded distribution in Section 4. However, when* $\mathcal{D}$ *is an unbounded distribution we cannot ignore events with a low probability and other techniques might be required to estimate the utilities. On a positive note, we still observe an increase in utility for the (unbounded) log-normal distribution in our empirical study (Section 4).*

4 EMPIRICAL OBSERVATIONS

We examine the effect of our constraints on naturally occurring distributions of utilities derived from scores in the *IIT-JEE 2009* dataset.[3] We consider with two disjoint groups of candidates G_a and G_b, representing male and female candidates respectively.[4] First, we analyze the distributions of the scores, $\mathcal{D}_a$ and $\mathcal{D}_b$, attained by G_a and G_b, and note that that the distributions of utilities of two groups are very similar in Section 4.1.1. In Section 4.2 we consider the situation in which these scores accurately capture a candidate's latent utility;[5] yet implicit bias against candidates in G_b, say during a job interview, affects the interpretation of the the score of these candidates. We simulate this implicit bias by shading the distribution of utilities for candidates in G_b by a constant multiplicative factor $\beta \in (0,1]$. We then measure the effectiveness of our constraints by comparing the latent utilities of:

(1) Cons: Our proposed ranking, which uses constraints to correct for implicit bias.
(2) Uncons: An unconstrained ranking.
(3) Opt: The optimal (unattainable) ranking that maximizes the (unobserved, due to implicit bias) utilities.

In Section 4.3, we consider the situation in which the scores themselves encode systematic biases, and consider the effectiveness of our constrained ranking as a potential intervention. In particular, we contrast our approach with a recent intervention used in IIT admissions, Supernumerary, which was created to increase the representation of women at IITs [5].

4.1 Dataset: JEE scores

Indian Institutes of Technology (IITs) are a group of 23 institutes of higher-education, which are, arguably, the most prestigious engineering schools in India.[6] Currently, undergraduate admissions into the IITs are decided solely on the basis of the scores attained in

[3] We observe similar results from the citation dataset from the *Semantic Scholar Open Research Corpus*, which we present in Section E.2 of the Supplementary Material.

[4] While there could be richer and non-binary gender categories, the above datasets code all of their entries as either male or female.

[5] This may not be the case as examined further in Section 4.3

[6] The number of IITs in 2009, the year the dataset is from, was 15.

the Joint Entrance Exam (JEE Advanced; known as IIT-JEE in 2009). IIT-JEE is conducted once every year, and only students that have graduated from high school in the previous two years are eligible. Out of the 468, 280, candidates who took IIT-JEE in 2011, only 9627 candidates (2%) were admitted to an IIT. In the same year, 108, 653 women (23.2% of the total) appeared in the exam, yet only 926 were admitted into an IIT (less than 10% of the 9627 admitted)[31].

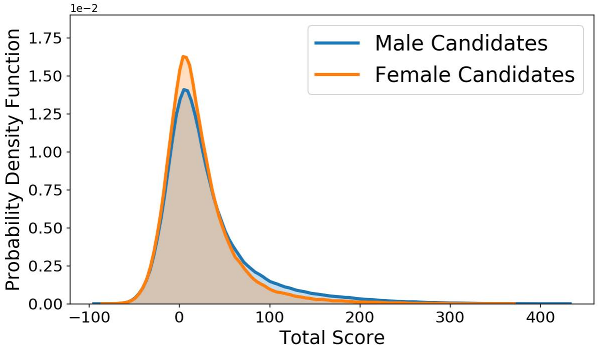

Figure 1: *Distributions of Scores in IIT-JEE 2009:* Distribution of total scores of all male and all female candidates. Men and women have similar distributions of total scores, with a total variation distance of $\Delta_{\text{TV}}(\mathcal{D}_a, \mathcal{D}_b) = 0.074$.

The dataset consists of scores of candidates from IIT-JEE 2009 which was released in response to a Right to Information application filed in June 2009 [35]. This dataset contains the scores of 384,977 students in each of the Math, Physics, and Chemistry sections of IIT-JEE 2009, along with the student's gender, given as a binary label (98,028 women and 286,942 men), their birth category (see [5]), and zip code. The candidates are scored on a scale from −35 to 160 points in all three sections, with an average total score of 28.36, a maximum score of 424 and a minimum score of −86. While the statistics of IIT JEE 2009 admissions are not available, if roughly the same number of students were admitted in 2009 as in 2011, then students with a score above 170 would have been admitted.

4.1.1 Distribution of scores across groups. We found that the Johnson's S_U-distribution gave a good fit for the distribution of scores. These fitted distributions of scores of women and men, $\mathcal{D}_b$ and $\mathcal{D}_a$, are depicted in Figure 1. The two distributions are very similar; their total-variation distance is $\Delta_{\text{TV}}(\mathcal{D}_a, \mathcal{D}_b) = 0.074$, i.e., the two distributions differ on less than 8% of their probability mass. However, the mean of men $\hat{\mu}_a = 30.79$ (standard deviation $\hat{\sigma}_a = 51.80$) is considerably higher than the mean of women $\hat{\mu}_b = 21.24$ (standard deviation $\hat{\sigma}_a = 39.27$).

4.2 Effectiveness of constraints

We now evaluate the effectiveness of the constraints as an intervention for implicit bias (see Section 2.2). For this evaluation, we assume that the JEE scores represent the true latent utility of a candidate, and we assume these scores are distributed for male and female students according to the fitted distributions $\mathcal{D}_a$ and $\mathcal{D}_b$ respectively. We then consider the case where k students, m_a from group G_a and m_b from group G_b apply to a job where the hiring manager has implicit bias β against group G_b. Here, Uncons would rank the candidates according to the biased utilities, Opt would rank the candidates according to their latent utilities (which is apriori impossible due to the implicit bias), and the proposed solution Cons would provide the optimal constrained ranking satisfying constraint parameter α using the fast-greedy algorithm described in [17]. More formally, the rankings are

$$\begin{aligned} \text{Uncons} &:= \operatorname{argmax}_x \mathcal{W}(x, v, \hat{w}), \\ \text{Cons} &:= \operatorname{argmax}_{x \in \mathcal{K}(L(\alpha))} \mathcal{W}(x, v, \hat{w}), \\ \text{Opt} &:= \operatorname{argmax}_x \mathcal{W}(x, v, w). \end{aligned}$$

We report the average latent utilities $U_{\mathcal{D},v}(\alpha, \beta)$, $U_{\mathcal{D},v}(0, \beta)$, and $U_{\mathcal{D},v}(0, 1)$ obtained by Cons, Uncons, and Opt respectively (see Equation (13) for the definition of $U_{\mathcal{D},v}(\cdot, \cdot)$).

4.2.1 Parameters. We let $m = m_a + m_b := 1000$, $k := 100$, $v_k := 1/\log(k+1)$ and vary the implicit bias parameter $\beta \in \{1/4, 1/2, 1\}$ and $m_b \in \{m/2, m/3, m/4\}$. In the JEE dataset, $m_b \approx m/4$ is a representative number. The position discount, $v_k = 1/\log(k+1)$ corresponds to DCG [30, 32, 55] which is popularly used in practice. We vary α from 0 to 0.5, i.e., from no constraint to the case where half of the candidates in the ranking must be from group G_b. We report the average latent utilities $U_{\mathcal{D},v}(\alpha, \beta)$, $U_{\mathcal{D},v}(0, \beta)$, and $U_{\mathcal{D},v}(0, 1)$; we take the average over 5000 trials in Table 1.

4.2.2 Results. We observe that the constraint intervention can significantly increase the latent utility of the ranking, even when, $\mathcal{D}_a$ and $\mathcal{D}_b$ are non-uniform, and are not exactly the same. The extent of the improvement depends both on the degree of implicit bias β, and the fraction of the underrepresented group m_a/m that appears in the dataset. As expected from Theorem 3.3, there exists an α for which the latent utility from the optimal ranking can be attained. More generally, we observe that the constraints increase the latent utility when α is such that it ensures no more than proportional representation – i.e., when the constraint ensures that the ranking reflects the group percentages in the underlying population, even when the bias parameter is small.

We conduct similar simulations without a discounting factor – i.e., with just selection with no ranking, and observe similar results in Section E.1 in the Supplementary Material. We also observe similar results on the Semantic Scholar Research Corpus, where the distributions are heavy tailed and hence constitute a very different class of utilities (see Section E.2 in the Supplementary Material). Overall, these findings suggest that the intervention is robust and across different latent utility distributions, population sizes, and extent of bias, and that the exact parameters need not be known to develop successful interventions.

4.3 IIT supernumerary seats for women

If we assume that the scores of the candidates are a true measure of their academic "potential", then any scheme which increases the number of underrepresented candidates admitted is bound to decrease the average "potential" of candidates at the institute. However, among candidates of equal "potential", those from an underprivileged group are known to perform poorer on standardized tests [54]. In India, fewer girls than boys attend primary school [3, 39], many of whom are forced to drop-out of schools to help with work at home or get married [59]. Therefore, we expect a female student who has the same score as a male student, say in

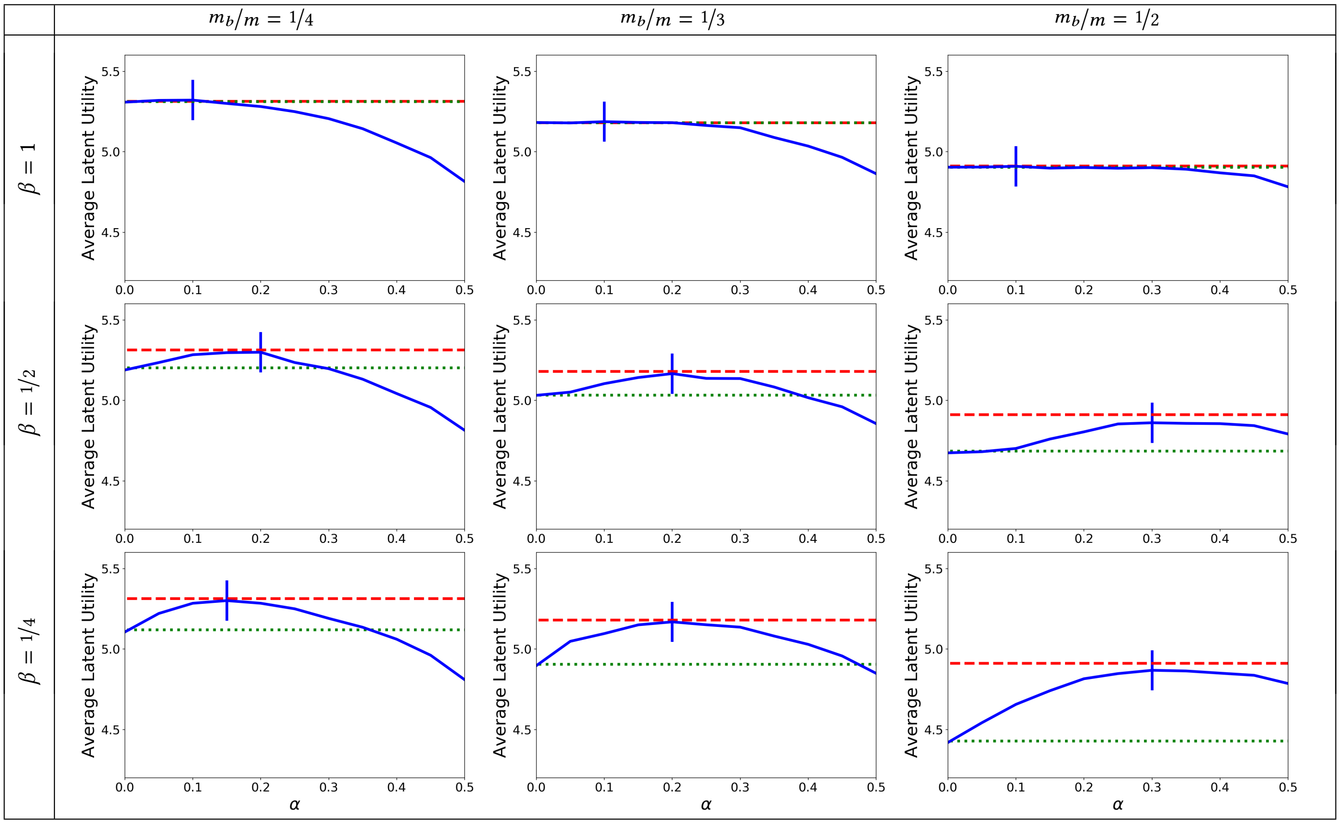

Table 1: *Empirical Results on IIT-JEE 2009 Dataset (With DCG):* We plot the latent utilities, $U_{\mathcal{D},v}(\alpha,\beta)$, $U_{\mathcal{D},v}(0,\beta)$, and $U_{\mathcal{D},v}(0,1)$ obtained by Cons, Uncons and Opt respectively (see Equation (13) for the definition of $U_{\mathcal{D},v}(\cdot,\cdot)$); we average over values over $5\cdot 10^3$ trials. Each plot represents an instance of the problem for a given value of implicit bias parameter β and the ratio of the size, m_b, of the underprivileged group, to the size m_a, of the privileged group. The bar represents the optimal constraint α: where we require the ranking to place at least $k\alpha$ candidates in the top k positions of the ranking for every position k. We note that our constraints attain close to the optimal latent utility for the right choice of α. More notably, they significantly outperform the unconstrained setting for a wide range of α, lending robustness to the approach. Even when there is no observed bias ($\beta = 1$), adding constraints does not appear to significantly impact the utility of the ranking unless the constraints are very strong.

IIT-JEE, to perform better than the boy if admitted. This is a societal bias, which, while different in nature than implicit bias, is another reason through which utilities can be systematically wrong against a particular group. Hence, the constraint approach presented is equally applicable as we illustrate in this section. In effect, it means that the scores are in fact biased, and the true latent utility of a candidate from G_b is in fact larger than what is reflected from their score.

To account for this and improve the representation of women in IITs, in 2018, additional "supernumerary" seats for women were introduced. This increases the capacity of all majors in all IITs by creating additional seats which are reserved for women, such that women compose at least 14% of each major, without decreasing the number of non-female candidates admitted. More formally, if the original capacity of a major at an IIT was C, and the average number of females admitted in this major in the past few years was n_f, then the scheme created x additional seats such that $n_f + x := 0.14(C+x)$. Where $(n_f + x)$ seats are reserved for females and the other $(C - n_f)$ seats are open to candidates of all genders. Eligible women are granted admission on reserved seat first, and only when all reserved seats are filled a women admitted on a gender neutral seat [5].

4.3.1 Setup. We assume the number of slots available in a given year for the IITs is $n := 10^4$.[7] We assume that the true latent utility of a candidate from group G_b is given by a shift $\gamma > 1$, such that if they attain a score s_f, then a true score would be $s'_f := (s_f + 105)\cdot\gamma - 105$.[8] In our simulations, we use $\gamma \approx 1.076$, which results in a shifted distribution $\mathcal{D}'_b$ with the same mean as $\mathcal{D}_a$.

Let the supernumerary scheme, Sup(α), admit $n_{\text{Sup}}(\alpha) \geq n$ candidates. We know that Sup always admits more candidates than our constraints approach, Cons(α), and the unconstrained approach Uncons, both of which choose n candidates. As such, it would be unfair to compare the average utility of the candidate selected by Sup with that of Cons(α) or Uncons. We define two more rankings $\overline{\text{Cons}}(\alpha)$ and $\overline{\text{Uncons}}$, which given an α select $n_{\text{Sup}}(\alpha)$ candidates and are otherwise equivalent to Cons(α) and Uncons.

[7] This number is close to the 9311 and 9576 students admitted into IITs in 2011 and 2012 which had the number of IITs as 2009.
[8] The shifts by 105 are to account fo the score range which can be negative.

We allocate n seats using the simple constraints scheme, $\textsc{Cons}$ and $\textsc{Uncons}$ and $n_{\textsc{Sup}}(\alpha) \geq n$ seats using $\textsc{Sup}(\alpha)$, $\overline{\textsc{Cons}}(\alpha)$, and $\overline{\textsc{Uncons}}$. and compare the average latent utilities of the candidates admitted by the scheme. Here, we define the average latent utility from scheme A which admits $n(A)$ candidates as

$$U(A) := \frac{1}{n(A)} \cdot \mathbb{E}_{w \leftarrow \mathcal{D}_a, \mathcal{D}'_b}[\mathcal{W}(\tilde{x}, v, w)]$$
$$\text{where } \tilde{x} := \max_{x \text{ satisfies } A} \mathcal{W}(\tilde{x}, v, \hat{w}).$$

We vary α from 0.1 to 0.25, i.e., from the fraction of women admitted if all candidates are admitted on basis of their scores, to the value which corresponds to proportional the representation based on the number of candidates appearing in IIT-JEE 2009. We report the average latent utilities in Figure 2.

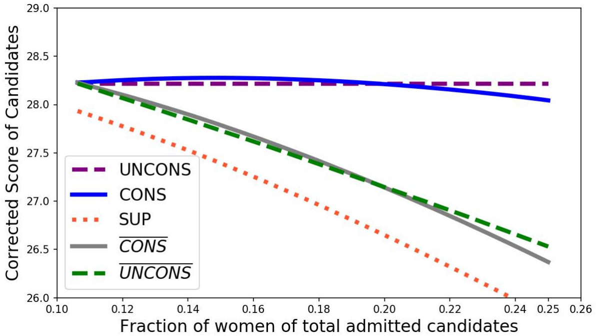

Figure 2: *Empirical Results Studying the Effect of Supernumerary Seats for Women:* We plot the average latent utilities of the ranking schemes we consider. The y-axis represents the utility per candidate , and the x-axis represents the lower-bound constraint α. Our constrained interventions outperform the unconstrained variants up to approximately 20% female seats, and always outperforms the existing supernumerary approach used in practice.

4.3.2 Results. We observe that the $\textsc{Sup}(\alpha)$ always has an lower average utility that $\overline{\textsc{Cons}}(\alpha)$ scheme. We note that the $\textsc{Sup}(\alpha)$ and the $\overline{\textsc{Cons}}(\alpha)$ always select the same number of candidates. The difference is that $\textsc{Sup}(\alpha)$ could place all the underprivileged candidates at the end of the ranking, while $\overline{\textsc{Cons}}(\alpha)$ places at least $k \cdot \alpha$ underprivileged candidates every k positions.

We find that for any $\alpha > 0.1$, $\textsc{Sup}(\alpha)$ decreases the average latent utility of the admitted candidates.

Finally, we observe that for a range of α (from 11% to 19.4%), $\textsc{Cons}$ increases the average latent utility of the admitted candidates over $\textsc{Uncons}$ and $\overline{\textsc{Cons}}(\alpha)$ increases the latent utility over $\overline{\textsc{Uncons}}$, i.e., for all $\alpha \in [0.11, 0.194]$,

$$1/n \cdot U(\textsc{Cons}(\alpha)) > 1/n \cdot U(\textsc{Uncons}),$$
$$1/n_{\textsc{Sup}}(\alpha) \cdot U(\overline{\textsc{Cons}}(\alpha)) > 1/n \cdot U(\overline{\textsc{Uncons}}).$$

The optimal constraint for $\textsc{Cons}(\alpha)$ is $\operatorname{argmax}_\alpha U(\textsc{Cons}(\alpha)) = 0.15$. Intuitively, $\textsc{Cons}(\alpha)$ can increase the average utility by swapping a male candidate $i \in G_a$ and female candidate $j \in G_b$, such that $\hat{w}_j < w_i < w_j$. Note that in $\textsc{Cons}(\alpha)$ additional candidates selected from G_b compete against the lowest scoring candidates from G_b in the ranking, instead of the average utility $U(\textsc{Cons}(\alpha))$ as in $\textsc{Sup}(\alpha)$.

Remark 4.1. *Since $\textsc{Sup}(\alpha)$ was not designed to optimize the average utility, it is not surprising that $\overline{\textsc{Cons}}(\alpha)$ outperforms $\textsc{Sup}(\alpha)$ in our experiment. However, the goal of this experiment is to study the effect our constraints approach against systematic biases different from implicit bias, by qualitatively comparing them an existing scheme ($\textsc{Sup}(\alpha)$).*

5 DISCUSSION AND LIMITATIONS

One could also consider other approaches to mitigate implicit bias. For instance, in a setting where an interviewer ranks the candidates, we could ask interviewers to self-correct (rescale) their implicit bias. However, anti-bias training, which aims to correct peoples' implicit bias, has been shown to have limited efficacy [43]. Thus, even with training, we do not have a guarantee that the interviewer can self-correct effectively. Adding interventions like the ones we consider do not require individuals to self-correct their perceptions.

Instead, we could ask interviewers to report their observed utilities and later rescale them. However, we may not have an accurate estimate of β. As the interventions we consider are independent of β (in Theorem 3.3), they can be applied in such a setting and would still recover the optimal utility (see Remark 3.4).

Furthermore, if interviewers are *explicitly* biased, they can give arbitrarily low scores to one group of candidates; this would make any rescaling attempt insufficient. By instead requiring a fixed number of candidates from each group, the interventions we consider are also robust against explicit bias, and perhaps this is why simple versions have been used in practice (e.g., Rooney rule [11, 49]).

We crucially note that any such intervention will only have a positive end effect if the goal is sincere; a hiring manager who is biased against a group of people can simply not hire a person from that group, regardless of what ranking they are given or what representation is enforced throughout the interview process. Interventions such as the one we describe here are a robust approach to correct for certain kinds of biases, but are only a single step in the process and alone cannot suffice. It would be important to evaluate this approach as part of the larger ecosystem in which it is used in order to adequately measure its efficacy in practice.

6 CONCLUSION

We consider a type of constraint-based interventions for re-ordering biased rankings derived from utilities with systematic biases against a particular socially salient group. The goal of these interventions is to recover, to the best of our ability, the unbiased ranking that would have resulted from the true latent (unbiased) utilities. We consider a theoretical model of implicit bias, and study the effect of such interventions on rankings in this setting.

We show that this family of constraint-based interventions are sufficient to mitigate implicit bias under this model; in other words, a ranking with the optimal latent utility can be recovered using this kind of intervention. However, the optimal parameters of the intervention depend on the specific utilities. Towards understanding this further, we make a natural assumption that the utilities are drawn from a fixed distribution, and show that simple constraints recover the optimal utility under this bias model. We focus on specific distributions, but believe that similar theoretical bounds would exist for other bounded distributions as discussed in Section 3.3. Rigorously analyzing necessary properties of distributions for which such bounds hold would be an interesting direction for future work.

This class of interventions is robust against *explicit* bias (see Section 5). Their robustness is further supported by our empirical findings, in which we find that there exist optimal constraints $\alpha^\star$ for which the optimal latent utility is almost attained in expectation. Importantly, we find that the interventions are near-optimal even when the distributions are unbounded (e.g., lognormal distributions; see Section E.2 in the Supplementary Material). Further, the interventions remain near-optimal even when the latent utility distribution of the underprivileged group is similar, but not identical, to that of the privileged group (see Section 4.2). More generally, we observe that the intervention improves the latent utility of the ranking for a wide range of α, with the highest improvement roughly centered around proportional representation. This gives an interesting rule of thumb, but more importantly shows the robustness of the method; without knowing the exact optimal α, one can still improve the ranking's latent utility significantly. From these observations, we expect this class of constraint-based intervention to be successful for a wide class of settings; exploring its limitations and developing clear guidelines on when and how to use the interventions in a particular use case would be an important avenue for further work.

In this work we pose the problem in terms of implicit biases. However, the source of the bias is not important; our results hold as long as there is systematic bias against one group as captured by our model in Section 2.2. We briefly discuss a different type of social bias in Section 4.3. However, we crucially note that any such intervention will only have a positive end effect if the goal is sincere (see Section 5).

Lastly, we note that while we phrase the majority of our results in light of rankings, they also have implications for the subset selection problem (where a group must be chosen, but need not be ordered) by taking the discounting factor $v_k = 1$ for all positions $k \in [n]$. This, in effect, eliminates the importance of the order and the total utility depends only on the set of people or items selected. In particular, our results answers a question asked in [33] on the efficacy of the ℓ-th order Rooney Rule when the distribution of utilities is uniform. We further report empirical results for subset selection, which follow the setup and conclusions discussed in Section E.1, in the Supplementary Material.

Hence, simple constraint-based interventions appear to be highly effective and robust for a wide variety of biases, distributions and settings; it is our hope that they be adopted and studied further as a simple yet powerful tool to help combat biases in a variety of applications.

7 PROOFS

7.1 Proof of Theorem 3.1

Theorem 7.1. **(Restatement of Theorem 3.1).** *Given a set of latent utilities $\{w_i\}_{i=1}^m$, there exists constraints $L(w) \in \mathbb{Z}_{\geq 0}^{n\times p}$, such that, for all implicit bias parameters $\{\beta_s\}_{s=1}^p \in (0,1)^p$, the optimal constrained ranking $\tilde{x} := \mathrm{argmax}_{x\in\mathcal{K}(L(w))}\mathcal{W}(x,v,\hat{w})$ satisfies*

$$\mathcal{W}(\tilde{x},v,w) = \max\nolimits_x \mathcal{W}(x,v,w). \tag{18}$$

Proof. Let $\pi : [m] \to [n+1]$ be the observed utility maximizing ranking and $\pi^\star : [m] \to [n+1]$, be the optimal *latent* utility maximizing ranking, where we define $\pi(i) := n+1$ and $\pi^\star(i) := n+1$, if item i was not ranked in the first n positions of π or $\pi^\star$. We claim that the following constraints are suitable for Theorem 3.1: For all $s \in [p]$ and $k \in [n]$:

$$L_{ks} := \sum\nolimits_{t\in G_s\,:\,\pi^\star(t)\leq k} 1, \tag{19}$$

We will prove that under these constraints, π and $\pi^\star$ are the same up to the groups of the items ranked at each position. The following lemma shows that they also have the same latent utility.

Lemma 7.2. *If for all $k \in [n]$, $T_{i_k} = T_{i_k^\star}$ where $i_k, i_k^\star$ are such that $i_k = \pi^{-1}(k)$ and $i_k^\star = (\pi^\star)^{-1}(k)$, i.e., π and $\pi^\star$ are the same up to the groups of the items ranked at each position, then π and $\pi^\star$ have the same latent utility.*

Proof. We show that the relative order of all items with different latent utilities is the same between the two rankings, proving that they have the same latent utility. Consider two items i_1 and i_2 in the same set of groups, i.e., such that $T_{i_1} = T_{i_2}$. We note that swapping their positions does not violate any new constraints. Further since, i_1 and i_2 have the same implicit bias, we have

$$\left(T_{i_1} = T_{i_2} \text{ and } w_{i_1} > w_{i_2}\right) \implies \hat{w}_{i_1} > \hat{w}_{i_2}. \tag{20}$$

Towards a contradiction, assume that π and $\pi^\star$ have different relative order of i_1 and i_2. Without loss of generality let $\pi^\star(i_1) < \pi^\star(i_2)$ and $\pi(i_1) > \pi(i_2)$. Since $\pi^\star$ is optimal we have $w_{i_1} \geq w_{i_2}$. If $w_{i_1} = w_{i_2}$, then these items do not change the latent utility between $\pi^\star$ and π. Let $w_{i_1} < w_{i_2}$, from Equation (20) we have that $\hat{w}_{i_1} < \hat{w}_{i_2}$. Therefore, we can swap the positions of i_1 and i_2 in π to gain the following observed utility

$$\begin{aligned}\hat{w}_{i_1}v_{\pi(i_2)} + \hat{w}_{i_2}v_{\pi(i_1)} - \hat{w}_{i_1}v_{\pi(i_1)} - \hat{w}_{i_2}v_{\pi(i_2)}\\ = \left(\hat{w}_{i_2} - \hat{w}_{i_1}\right)\cdot\left(v_{\pi(i_1)} - v_{\pi(i_2)}\right) > 0.\end{aligned}$$

This contradicts the fact that π has the optimal observed utility. □

It remains to prove that with these constraints, π is the same as $\pi^\star$ up to the groups of the items. This proof is by induction.
Inductive hypothesis: Let the two rankings agree on the first $(k-1)$ positions (up to the groups of the item ranked at each position).
Base case: The base case for $k = 1$ is trivially satisfied.
Towards, a contradiction assume that the items on the k-th positions of π and $\pi^\star$ are i and $i^\star \in [m]$, such that $T_i \neq T_{i^\star}$. Since $\pi^\star$ ranks i after $i^\star$ it follows that

$$w_i \leq w_{i^\star} \tag{21}$$

Case A: $T_i \subsetneq T_{i^\star}$: We claim that this case is not possible. Consider any group G_s, for $s \in T_{i^\star}\backslash T_i$. Since the rankings agree on the first $(k-1)$ positions we have for all $j \in [k-1]$

$$\sum\nolimits_{t\in G_s\,:\,\pi(t)\leq j} 1 = \sum\nolimits_{t\in G_s\,:\,\pi^\star(t)\leq j} 1. \tag{22}$$

Since $\pi(i) = k$ we have

$$\begin{aligned}\sum\nolimits_{t\in G_s:\,\pi(t)\le k} 1 &= \sum\nolimits_{t\in G_s:\,\pi(t)<k} 1 + \mathbb{I}[i\in G_s]\\ &\overset{(22)}{=} \sum\nolimits_{t\in G_s:\,\pi^\star(t)<k} 1 + \mathbb{I}[i\in G_s]\\ &< \sum\nolimits_{t\in G_s:\,\pi^\star(t)<k} 1 + \mathbb{I}[i\in G_s] + \mathbb{I}[i^\star\in G_s] \quad \text{(Using } i^\star\in G_s)\\ &= \sum\nolimits_{t\in G_s:\,\pi^\star(t)\le k} 1 + \mathbb{I}[i\in G_s]\\ &= \sum\nolimits_{t\in G_s:\,\pi^\star(t)\le k} 1 \overset{(19)}{=} L_{ks}. \quad \text{(Using } i\notin G_s)\end{aligned}$$

Therefore, in this case π violates the constraint at position k.

Case B: $T_i \supsetneq T_{i^\star}$**:** In this case we have

$$\begin{aligned}\hat{w}_i &= \Big(\prod\nolimits_{s\in[p]:\,G_s\ni i}\beta_s\Big)\cdot w_i = \Big(\prod\nolimits_{s\in[p]}\beta_s^{\mathbb{I}[i\in G_s]}\Big)w_i\\ &= \Big(\prod_{s\in[p]}\beta_s^{\mathbb{I}[i^\star\in G_s]}\Big)\Big(\prod_{s\in[p]}\beta_s^{\mathbb{I}[i\in G_s]\cdot\mathbb{I}[i^\star\notin G_s]}\Big)w_i \quad \text{(Using } T_i\supsetneq T_{i^\star})\\ &< \Big(\prod\nolimits_{s\in[p]}\beta_s^{\mathbb{I}[i^\star\in G_s]}\Big)\cdot w_i \quad \text{(For all } s\in[p],\ \beta_s\in(0,1))\\ &\overset{(21)}{<} \Big(\prod\nolimits_{s\in[p]}\beta_s^{\mathbb{I}[i^\star\in G_s]}\Big)\cdot w_{i^\star} < \hat{w}_{i^\star}. \qquad (23)\end{aligned}$$

Consider the ranking $\hat{\pi}$ formed by swapping the position of i and $i^\star$ in π. The change in observed utility between $\hat{\pi}$ and π is

$$\begin{aligned}&\hat{w}_i v_{\pi(i^\star)} + \hat{w}_{i^\star} v_{\pi(i)} - \hat{w}_i v_{\pi(i)} - \hat{w}_{i^\star} v_{\pi(i^\star)}\\ &\qquad = \big(\hat{w}_{i^\star} - \hat{w}_i\big)\cdot\big(v_{\pi(i)} - v_{\pi(i^\star)}\big) \overset{(23)}{>} 0.\end{aligned}$$

Therefore, $\hat{\pi}$ has higher observed utility than π. Therefore, if $\hat{\pi}$ satisfies the constraints, then this contradicts the optimality of π, and we are done. $\hat{\pi}$ can only violate the constraints in positions $j \in [k, \dots, \pi(i^\star)]$ for groups $T_i \backslash T_{i^\star}$. Since the number of items selected from the set of groups $T_i \backslash T_{i^\star}$ only increases. It follows that if π satisfies the lower bound constraints on positions $j \in [k, \dots, \pi(i^\star)]$, then so does $\hat{\pi}$. Since we assume π to be feasible, we are done. □

7.2 Proof of Theorem 3.2

Theorem 7.3. **(Restatement of Theorem 3.2).** *Let $\mathcal{D}$ be a continuous distribution, $\ell \le m_b$, and $0 < k < \min(m_a, m_b)$ be a position, then*

$$\forall\ \delta \ge 2 \qquad \Pr[\ N_{kb} \le \mathbb{E}[N_{kb}] - \delta\ ] \le e^{-\frac{(2\delta^2-1)}{k}}, \qquad (24)$$

$$\mathbb{E}[N_{kb}] = k \cdot {}^{m_b}/_{(m_a+m_b)}, \qquad (25)$$

$$\mathbb{E}[P_\ell] = \ell \cdot \big(1 + {}^{m_a}/_{(m_b+1)}\big). \qquad (26)$$

Equivalent Simple Model. Since the items have the same distribution $\mathcal{D}$ of latent utility, the event: $i \in G_a$ (or equivalently $i \in G_b$), is independent of the event: $w_i = z$ for some $z \in \text{supp}(\mathcal{D})$. Therefore, the latent utility maximizing ranking is independent of the groups assigned to each item.

It follows that the following is an equivalent model of generating the unbiased ranking: First, we draw $m_a + m_b$ latent utilities and order them in a non-increasing order (randomly breaking ties). Then, we choose m_b items uniformly without replacement and assign them to G_b, and assign the others to G_a. Let a item $i \in G_b$ be a blue ball and $i \in G_a$ be a red ball, then distributions of positions of the balls is equivalent to: Given $m_a + m_b$ numbered balls, we pick m_b of them uniformly without replacement and color them blue, and color the rest of the balls red.

Proof. Using the above model, we first find that P_ℓ has a negative-hypergeometric distribution and use it to calculate, $\mathbb{E}[P_\ell]$, then we find that N_{kb} is a hypergeometric distribution, use this fact to calculate $\mathbb{E}[N_{kb}]$ and show N_{kb} is concentrated around its mean.

Expectation of P_ℓ. The total ways of choosing m_b blue balls is $\binom{m_a+m_b}{m_b}$. Given $P_\ell = k$, the number of ways of choosing the $\ell - 1$ blue balls before it is $\binom{k-1}{\ell-1}$ and the number of ways of choosing the $b - \ell$ blue balls after it is $\binom{m_a+m_b-k}{b-\ell}$. Therefore by expressing the probability as the ratio of favorable and total outcomes we have

$$\Pr[P_\ell = k] = \tbinom{k-1}{\ell-1}\tbinom{m_a+m_b-k}{b-\ell}/\tbinom{m_a+m_b}{m_b}. \qquad (27)$$

Rewriting it using the fact that $\binom{m_a}{m_b} = \binom{m_a}{m_a-m_b}$ we get

$$\Pr[P_\ell = k] = \frac{\binom{k-1}{k-\ell}\binom{m_a+m_b-k}{m_a-b+\ell}}{\binom{m_a+m_b}{m_a}} = \frac{\binom{(k-\ell)+(\ell-1)}{k-\ell}\binom{m_a+m_b-(k-\ell)-\ell}{m_a-(x-\ell)}}{\binom{m_a+m_b}{m_a}}.$$

Comparing this with the density function of the Negative hypergeometric distribution (28) it is easy to observe that $P_\ell - \ell$ is a negative hypergeometric variable.

Given numbers N, K, r, the negative hypergeometric random variable, NG, has the following distribution, for all $k \in [K]$

$$\Pr[NG = k] := \tbinom{k+r-1}{k}\tbinom{N-r-k}{K-k}/\tbinom{N}{K}, \qquad (28)$$

From the expectation of a negative-hypergeometric variable we have

$$\mathbb{E}[P_\ell - \ell] = \ell \cdot \frac{m_a}{m_b + 1} \implies \mathbb{E}[P_\ell] \quad = \ell \cdot \frac{(m_a + m_b + 1)}{m_b + 1}. \qquad (29)$$

Expectation and Concentration of N_{kb}. Given $N_{kb} = j$, the number of ways of coloring j out of k balls before it is $\binom{k}{j}$, and the number of ways of coloring $b - j$ balls after it is $\binom{m_a+m_b-k}{b-\ell}$. Therefore it follows that:

$$\Pr[\ N_{kb} = j\ ] = \tbinom{k}{j}\tbinom{m_a+m_b-k}{b-j}/\tbinom{m_a+m_b}{m_b}. \qquad (30)$$

Given numbers N, K, n, for an hypergeometric random variable, HG, we have: $\forall \max(0, n + K - N) \le k \le \min(K, n)$

$$\Pr[HG = k] := \tbinom{K}{k}\tbinom{N-K}{n-k}/\tbinom{N}{n}. \qquad (31)$$

By comparing with Equation (31), we can observe that N_{kb} is a hypergeometric random variable. From well known properties of the hypergeometric distribution we have that

$$\mathbb{E}[\ N_{kb}\ ] = {}^{kb}/_{(m_a+m_b)} \qquad (32)$$

$$\Pr\left[N_{kb} \ge {}^{kb}/_{(m_a+m_b)} + \delta\right] \overset{[29]}{\le} e^{-2(\delta^2-1)\gamma} \le e^{-\frac{2(\delta^2-1)}{k+1}} \qquad (33)$$

$$\Pr\left[N_{kb} \le {}^{kb}/_{(m_a+m_b)} - \delta\right] \overset{[29]}{\le} e^{-2(\delta^2-1)\gamma} \le e^{-\frac{2(\delta^2-1)}{k+1}}. \qquad (34)$$

where $\gamma := \max\big(\frac{1}{m_a+1} + \frac{1}{m_b+1}, \frac{1}{k+1} + \frac{1}{m_a+m_b-n+1}\big) \ge \frac{1}{k+1}$.

□

Acknowledgment. *Part of this research was supported by an AWS MLRA 2019 Award.*

REFERENCES

[1] ACM. 2017. Statement on Algorithmic Transparency and Accountability. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

[2] Social Security Administration. 2018. Beyond the Top 1000 Names. https://www.ssa.gov/oact/babynames/limits.html.

[3] Harold Alderman and Elizabeth M King. 1998. Gender differences in parental investment in education. *Structural Change and Economic Dynamics* 9, 4 (1998), 453–468.

[4] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing Fair Ranking Schemes. In *SIGMOD Conference*. ACM, 1259–1276.

[5] Surender Baswana, P. P. Chakrabarti, Yashodhan Kanoria, Utkarsh Patange, and Sharat Chandran. 2019. Joint Seat Allocation 2018: An algorithmic perspective. *CoRR* abs/1904.06698 (2019). arXiv:1904.06698 http://arxiv.org/abs/1904.06698

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

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

[8] Miranda Bogen and Aaron Rieke. 2018. Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias. https://www.upturn.org/reports/2018/hiring-algorithms/.

[9] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 2013. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

[10] Carlos Castillo. 2019. Fairness and Transparency in Ranking. In *ACM SIGIR Forum*, Vol. 52. ACM, 64–71.

[11] Marilyn Cavicchia. 2017. How to fight implicit bias? With conscious thought, diversity expert tells NABE. (June 2017). https://www.americanbar.org/groups/bar_services/publications/bar_leader/2015-16/september-october/how-fight-implicit-bias-conscious-thought-diversity-expert-tells-nabe/

[12] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *FAT*. ACM, 319–328.

[13] L. Elisa Celis, Lingxiao Huang, and Nisheeth K. Vishnoi. 2018. Multiwinner Voting with Fairness Constraints. In *IJCAI*. ijcai.org, 144–151.

[14] L. Elisa Celis, Sayash Kapoor, Farnood Salehi, and Nisheeth K. Vishnoi. 2019. Controlling Polarization in Personalization: An Algorithmic Framework. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 160–169. https://doi.org/10.1145/3287560.3287601

[15] L. Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. 2018. Fair and Diverse DPP-Based Data Summarization. In *ICML (Proceedings of Machine Learning Research)*, Vol. 80. PMLR, 715–724.

[16] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2019. Toward Controlling Discrimination in Online Ad Auctions. In *ICML (Proceedings of Machine Learning Research)*, Vol. 97. PMLR, 4456–4465.

[17] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *ICALP (LIPIcs)*, Vol. 107. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 28:1–28:15.

[18] Brian W Collins. 2007. Tackling unconscious bias in hiring practices: The plight of the Rooney rule. *NYUL Rev.* 82 (2007), 870.

[19] Joshua Correll, Bernadette Park, Charles M. Judd, and Bernd Wittenbrink. 2007. The influence of stereotypes on decisions to shoot. *European Journal of Social Psychology* 37, 6 (2007), 1102–1117.

[20] Jeffrey Dastin. 2019. Amazon scraps secret AI recruiting tool that showed bias against women. https://reut.rs/2N1dzRJ.

[21] Jennifer L. Eberhardt and Sandy Banks. 2019. Implicit bias puts lives in jeopardy. Can mandatory training reduce the risk? https://www.latimes.com/opinion/op-ed/la-oe-eberhardt-banks-implicit-bias-training-20190712-story.html.

[22] Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521. https://doi.org/10.1073/pnas.1419828112 arXiv:http://www.pnas.org/content/112/33/E4512.full.pdf

[23] Facebook. [n.d.]. Managing Unconscious Bias. https://managingbias.fb.com.

[24] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. *CoRR* abs/1905.01989 (2019).

[25] Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review* 102, 1 (1995), 4.

[26] Anthony G Greenwald and Linda Hamilton Krieger. 2006. Implicit bias: Scientific foundations. *California Law Review* 94, 4 (2006), 945–967.

[27] The White House. 2015. Fact Sheet: President Obama Announces New Commitments from Investors, Companies, Universities, and Cities to Advance Inclusive Entrepreneurship at First-Ever White House Demo Day. (August 2015). https://obamawhitehouse.archives.gov/the-press-office/2015/08/04/fact-sheet-president-obama-announces-new-commitments-investors-companies

[28] Lingxiao Huang, Shaofeng H.-C. Jiang, and Nisheeth K. Vishnoi. 2019. Coresets for Clustering with Fairness Constraints. In *NIPS*.

[29] Don Hush and Clint Scovel. 2005. Concentration of the hypergeometric distribution. *Statistics & probability letters* 75, 2 (2005), 127–132.

[30] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[31] JEE Team 2011 – IIT Kanpur. 2011. JEE-2011 Report. http://bit.do/e5iEZ.

[32] Evangelos Kanoulas and Javed A Aslam. 2009. Empirical justification of the gain and discount function for nDCG. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 611–620.

[33] Jon M. Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. In *ITCS (LIPIcs)*, Vol. 94. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 33:1–33:17.

[34] Caitlin Kuhlman, MaryAnn VanValkenburg, and Elke Rundensteiner. 2019. FARE: Diagnostics for Fair Ranking using Pairwise Error Metrics. (2019).

[35] Mr. Rajeev Kumar. 2009. RTI Complaint. Decision No. CIC/SG/C/2009/001088/5392, Complaint No. CIC/SG/C/2009/001088.

[36] Karen S Lyness and Madeline E Heilman. 2006. When fit is fundamental: performance evaluations and promotions of upper-level female and male managers. *Journal of Applied Psychology* 91, 4 (2006), 777.

[37] Kay Manning. 2018. As Starbucks gears up for training, here's why 'implicit bias' can be good, bad or very bad. https://www.chicagotribune.com/lifestyles/sc-fam-implicit-bias-0529-story.html.

[38] R McGregor-Smith. 2017. Race in the Workplace: The Mcgregor-Smith Review. (2017). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/594336/race-in-workplace-mcgregor-smith-review.pdf.

[39] Government of India Ministry of Home Affairs. 2011. CensusInfo India 2011: Final Population Totals. http://censusindia.gov.in/2011census/censusinfodashboard/index.html.

[40] Corinne A Moss-Racusin, John F Dovidio, Victoria L Brescoll, Mark J Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences* 109, 41 (2012), 16474–16479.

[41] Cecilia Munoz, Megan Smith, and D. J. Patil. 2016. Big data: A report on algorithmic systems, opportunity, and civil rights. *Executive Office of the President. The White House* (2016).

[42] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2019. Pairwise Fairness for Ranking and Regression. arXiv:cs.LG/1906.05330

[43] Mike Noon. 2018. Pointless diversity training: unconscious bias, new racism and agency. *Work, Employment and Society* 32, 1 (2018), 198–209.

[44] Jason A Okonofua and Jennifer L Eberhardt. 2015. Two strikes: Race and the disciplining of young students. *Psychological science* 26, 5 (2015), 617–624.

[45] Christina Passariello. 2016. Tech Firms Borrow Football Play to Increase Hiring of Women. (September 2016). https://www.wsj.com/articles/tech-firms-borrow-football-play-to-increase-hiring-of-women-1474963562

[46] B. Keith Payne, Heidi A. Vuletich, and Jazmin L. Brown-Iannuzzi. 2019. Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences* 116, 24 (2019), 11693–11698. https://doi.org/10.1073/pnas.1818816116 arXiv:https://www.pnas.org/content/116/24/11693.full.pdf

[47] Melody S Sadler, Joshua Correll, Bernadette Park, and Charles M Judd. 2012. The world is not black and white: Racial bias in the decision to shoot in a multiethnic context. *Journal of Social Issues* 68, 2 (2012), 286–313.

[48] Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. *CoRR* abs/1901.10437 (2019).

[49] Deepa Seetharaman. 2015. Facebook Is Testing the 'Rooney Rule' Approach to Hiring. *The Wall Street Journal* (June 2015). https://blogs.wsj.com/digits/2015/06/17/facebook-testing-rooney-rule-approach-to-hiring/

[50] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD*. ACM, 2219–2228.

[51] Eric Luis Uhlmann and Geoffrey L Cohen. 2005. Constructed criteria: Redefining merit to justify discrimination. *Psychological Science* 16, 6 (2005), 474–480.

[52] Linda Van den Bergh, Eddie Denessen, Lisette Hornstra, Marinus Voeten, and Rob W Holland. 2010. The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal* 47, 2 (2010), 497–527.

[53] Joseph Walker. 2012. Meet the New Boss: Big Data. https://www.wsj.com/articles/SB10000872396390443890304578006252019616768.

[54] Gregory M Walton and Steven J Spencer. 2009. Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science* 20, 9 (2009), 1132–1139.

[55] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, Vol. 8. 6.

[56] Christine Wenneras and Agnes Wold. 2001. Nepotism and sexism in peer-review. *Women, sience and technology: A reader in feminist science studies* (2001), 46–52.

[57] Bernard E Whitley Jr and Mary E Kite. 2016. *Psychology of prejudice and discrimination*. Routledge.

[58] Joan C Williams. 2014. Double jeopardy? An empirical study with implications for the debates over implicit bias and intersectionality. *Harvard Journal of Law &*

Gender 37 (2014), 185.

[59] Rachel Williams. 2013. Why girls in India are still missing out on the education they need. https://www.theguardian.com/education/2013/mar/11/indian-children-education-opportunities.

[60] Kathleen Woodhouse. 2017. Implicit Bias – Is It Really? http://bit.do/e5iFn.

[61] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*. ACM, 22:1–22:6. https://doi.org/10.1145/3085504.3085526

[62] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo A. Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *CIKM*. ACM, 1569–1578.

[63] Colin A Zestcott, Irene V Blair, and Jeff Stone. 2016. Examining the presence, consequences, and reduction of implicit bias in health care: A narrative review. *Group Processes & Intergroup Relations* 19, 4 (2016), 528–542.