# Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints

Anay Mehrotra
Yale University
USA

Bary S. R. Pradelski
National Centre for Scientific Research (CNRS)
France

Nisheeth K. Vishnoi
Yale University
USA

## ABSTRACT

In selection processes such as hiring, promotion, and college admissions, implicit bias toward socially-salient attributes such as race, gender, or sexual orientation of candidates is known to produce persistent inequality and reduce aggregate utility for the decision maker. Interventions such as the Rooney Rule and its generalizations, which require the decision maker to select at least a specified number of individuals from each affected group, have been proposed to mitigate the adverse effects of implicit bias in selection. Recent works have established that such lower-bound constraints can be very effective in improving aggregate utility in the case when each individual belongs to at most one affected group. However, in several settings, individuals may belong to multiple affected groups and, consequently, face more extreme implicit bias due to this *intersectionality*. We consider independently drawn utilities and show that, in the intersectional case, the aforementioned non-intersectional constraints can only recover part of the total utility achievable in the absence of implicit bias. On the other hand, we show that if one includes appropriate lower-bound constraints on the intersections, almost all the utility achievable in the absence of implicit bias can be recovered. Thus, intersectional constraints can offer a significant advantage over a reductionist dimension-by-dimension non-intersectional approach to reducing inequality.

## KEYWORDS

Implicit bias, selection, Intersectionality, Intersectional biases, Rooney Rule

**ACM Reference Format:**
Anay Mehrotra, Bary S. R. Pradelski, and Nisheeth K. Vishnoi. 2022. Selection in the Presence of Implicit Bias: The Advantage of Intersectional Constraints. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3531146.3533124

## 1 INTRODUCTION

*Implicit bias* is the unconscious association, belief, attitude, skewed observation, or lack of awareness toward any socially-salient group, which may lead to systematic disadvantages for particular – often underprivileged – groups [45, 46, 52, 53]. The negative impact of implicit bias on certain groups of the population is well documented in many societal contexts [1, 64], including hiring [7, 63, 71, 87], university admissions [50, 68], and healthcare [20, 43, 51]. Instances of implicit bias in hiring and other selection processes include higher salaries for men than women despite the same qualifications [63], biased peer-review of fellowship applications against women [84], and stricter promotion standards for women in managerial positions [59]. Affected candidates can also face biases before participating in selection processes: For instance, they can face implicit bias in the form of lower teacher expectations [41], or harsher grading policies [79], which may further hurt their future prospects in hiring or promotion [65]. Implicit bias not only has adverse effects on individuals, but also on decision makers who may hire/promote less qualified candidates. Moreover, such biases can also affect downstream algorithms and policies, either through biased human decisions or through past-data used to inform these decisions, giving rise to biases that affect different groups differently [10, 29, 83].

Policy makers, private entities, and researchers have introduced a host of measures to counter adverse effects of implicit bias: affirmative action policies which increase representation of affected groups [9, 49, 66, 75, 78], structured interviews which reduce the scope for bias in evaluation criteria [5, 39, 70, 84], and anonymized evaluations that blind decision makers to the socially-salient attributes of applicants [42]. Significant effort has also been devoted to reduce implicit bias itself: training that exposes individuals to counter-stereotypical evidence opposing their implicit beliefs [38, 60, 86], enhanced accountability which enables enforcement of other interventions [13, 56, 58], and information campaigns that increase awareness about implicit biases [62].

Of interest here are affirmative action policies that introduce lower-bound constraints for groups adversely affected by implicit bias. A popular instantiation of this strategy is the Rooney rule, which requires the decision maker to select at least one individual from the affected group for interview. The hope is that during the interview, the decision maker will assess the "true" value of the individual [12] and this interaction will reduce their implicit bias [28]. In addition, variants of the Rooney Rule have also been used for the final stage of a selection process such as that for board membership or highly-priced entry jobs to directly counter the effects of implicit bias [18, 48, 66, 75].

Recently, some works have analyzed the effectiveness of Rooney Rule type constraints for selection and ranking processes [19, 35, 55]. In particular, [55] study the effectiveness of the Rooney Rule for selection in the presence of implicit bias when there is a single affected group. Here, there are $m$ *individuals*, where each individual $i \in \{1, 2, \ldots, m\}$ has a non-negative *latent utility* $w_i \geq 0$, that is the *value* it adds to the selection, and an *observed utility* $\widehat{w}_i \leq w_i$,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*FAccT '22, June 21–24, 2022, Seoul, Republic of Korea*
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533124

that is the decision maker's possibly biased estimate of $w_i$. The decision maker selects $n$ individuals with the maximum sum of observed utilities. [55] study a model where implicit bias acts via a multiplicative factor $0 < \beta \leq 1$: the observed utility of an individual $i$ belonging to the affected group is $\widehat{w}_i = \beta \cdot w_i$ while that of an unaffected individual $j$ is $\widehat{w}_j = w_j$. They argue that this model is a reasonable approximation of the empirical findings of [84], who find that in peer-reviewed evaluations for fellowships women's score were systematically scaled down compared to men with similar productivity. [55] study conditions on the parameters $n$, $m$, $\beta$, and the distribution of latent utility $w_i$, where the Rooney Rule increases the total latent utility of the selection. Under the same implicit bias model, [19] study a generalization of the Rooney Rule, where the decision maker is constrained to select at least $L \geq 1$ individuals from the affected group. [19] show that, for a single affected group, there is an $L$ for which the decision maker, constrained by the generalized Rooney Rule, achieves near-optimal latent utility.

*Intersectionality* posits that one needs to take into account the interconnected nature of the multiple socially-salient attributes, as opposed to viewing these attributes through a reductionist lens, that is, dimension-by-dimension [27]. Intersectionality can lead to the creation of overlapping and interdependent systems of discrimination or disadvantage, and there is a rich literature in social sciences and law that studies it [3, 14, 17, 23–27, 34, 54, 69, 76, 85]. Intersectional implicit bias also arises in selection processes. For instance, [84] find significantly lower scores for women unaffiliated with the evaluation committee compared to other women and to men unaffiliated with the committee in peer-reviewed applications for a fellowship in Sweden. Thus, neither gender nor affiliation alone explain the bias faced by individuals, and to understand this bias, a combination of the two attributes must be considered. Recently, intersectional bias has also been observed in the outputs of algorithms. For instance, [16] audit commercial image-based gender classifiers and find intersectional bias against Black women, and [80] report intersectional bias in contextualized word representations.

However, the intersectional nature of social groupings and, thus, biases has been largely overlooked when designing interventions to reduce or counter implicit bias. Further, data reporting, such as that by the U.S. Census Bureau, is mostly dimension-by-dimension (for example, by race or by gender) and omits intersectional data. Since this data is used to inform policies, it has inevitably led to policies that only focus on reducing inequality along one identity dimension at a time, as highlighted in a report by the European Union [77]. In other words, existing data reporting and, consequently, policies are *non-intersectional*. They specify lower-bound constraints on each affected group, but not on their intersections. But it is natural to expect that the individuals at the intersection of multiple affected groups face higher and, possibly, different implicit bias [15, 54, 84, 85]. In fact, as argued by [54], intersectional bias can be significantly higher and often compound, or multiply, the biases faced by individuals in single affected groups. Thus, the following question arises and is studied in this paper.

*Are non-intersectional constraints sufficient to recover the entire latent utility with intersecting affected groups or does one need to specify constraints across all intersections to achieve this?*

## 1.1 Our Contributions

We consider the effectiveness of lower-bound constraints on selection processes in the presence of intersectional implicit bias. To capture the effect of intersectionality on implicit bias, we consider an extension of the aforementioned model of [55] due to [19]: each individual may belong to zero, one, or more of the $p$ affected groups (such as the groups of all women or all Black people) (Section 2.2). For each group $\ell \in \{1, 2, \ldots, p\}$, there is an implicit bias parameter $0 < \beta_\ell \leq 1$ and the implicit bias experienced by an individual is the product of parameters of each group they belong to.

We compare non-intersectional and intersectional lower-bound constraints when the latent utilities are independently and identically distributed. Non-intersectional constraints specify the minimum number of individuals to be selected from each affected group (e.g., the groups of all women or all Black people). They do not specify the minimum number of individuals to be selected from a given intersection (e.g., the groups of all Black women, all non-Black women, all Black non-women, or all non-Black non-women). For each of the intersectional groups, intersectional constraints specify the minimum number of individuals to be selected from this intersection. To compare the relative efficacy of constraints, we consider a *utility ratio*, defined as the expected value of the ratio of the latent utility achieved under the constraint to the latent utility absent any implicit bias (Section 2.4). By definition, the utility ratio is a number between 0 and 1 and the goal of the policy maker is design interventions such that the corresponding utility ratio is 1.

We show that under general conditions on the distribution of latent utilities, no matter which *non-intersectional constraints* are deployed, the utility ratio is strictly less than 1 (Theorems 3.1 and 3.4). In particular, our result applies to distributions such as uniform, truncated normal, and truncated power-law distributions. Moreover, our result gives a quantitative bound on the maximum of utility ratio: it is at most $1-\phi$, where $\phi$ is positive and depends only on the implicit bias parameters and generic parameters of the distribution family, and independent of the number of candidates $m$. Concretely, when the utilities are uniformly distributed on the interval $[0, 1]$, there is a family of instances such that the maximum utility ratio achievable using only non-intersectional constraints can be as low as $8/9$ (Proposition 3.3). Further, we show that this result also holds for generalizations of the implicit bias model where, for instance, the implicit bias experienced by individuals in multiple groups is different than the product of the implicit bias parameters of the groups they belong to (Theorem 3.4). Thus, these results imply that, unlike the setting of a single affected group studied in [19, 35, 55], non-intersectional constraints may be insufficient to completely mitigate the effects of implicit bias in the presence of intersections.

On the positive side, we show that there are *intersectional* lower-bound constraints that, for any amount of implicit bias, recover utility ratio arbitrarily close to 1 (Theorem 3.2). This result extends for all continuous distributions of utility and for the generalizations of the implicit bias model considered above (Corollary 3.5). We show that these intersectional lower-bound constraints end up being just a function of the sizes of intersections and do not depend on the amount of implicit bias $0 < \beta_1, \ldots, \beta_p \leq 1$ or the specific utility distribution. In fact, the constraints require at least a near-proportional number of individuals from each intersection and,

hence, they can be employed in practice where the implicit bias parameters $\beta_1, \ldots, \beta_p$ and the distribution of utility are not known and can vary across contexts or over time. Thus, a policy maker may choose intersectional constraints in order to obtain a utility ratio arbitrarily close to 1.

Overall, our results imply that the advantage of intersectional constraints can be substantial and a reductionist dimension-by-dimension approach is not sufficient to mitigate the adverse effects of implicit bias. They provide a utilitarian reason for policy makers to choose intersectional constraints over non-intersectional constraints.

## 1.2 Related Work

*Implicit bias and empirics.* There are several theories about how implicit bias arises; e.g., [2, 4, 40, 44, 47, 61, 67, 81]. Specific examples include, [81] who propose that humans unconsciously use heuristics to overcome their limited computing ability. Such heuristics can take the form of stereotypes where one divides individuals into groups and, then, extrapolates the characteristics of specific individuals from the characteristics associated with their group(s) [2, 40]. Another theory suggests that using stereotypes was evolutionarily advantageous [47, 57]. One reason, according to [57], is that undervaluing the utility of unknown out-group individuals and, hence, avoiding contact with them reduced the risk of contracting new diseases. Apart from these theories, it has also been suggested that implicit bias "is a trace of [the individual's] past experience" [44] and that prior, explicit, racism has been channeled into implicit bias [61]. Regardless of the cause of implicit bias, studies identifying the adverse effects of implicit bias are abundant: from police shootings [73], promotion and hiring decisions [8, 59], education [63, 82], to peer-review [84].

*Models of implicit bias and decision-making in the presence of implicit bias.* A growing literature is studying decision making in the presence of implicit bias, ranging from works on set selection [35, 37, 55], ranking [19], to classification [11]. Among these, works on the set selection and ranking problems are directly related to our work. [55] introduce a mathematical model of implicit bias for a single affected group and study when the Rooney Rule increases the total latent utility of the selection. Unlike them, we consider multiple and intersectional affected groups and also consider generalizations of the Rooney Rule. [19] study the ranking problem, where the selected individuals also need to be ordered. Specializing their work to set selection: they consider the setting with a single affected group where the decision maker must select at least $L \geq 1$ individuals from the affected group and show that there are constraints which achieve near-optimal latent utility in expectation. [19] also extended the model of implicit bias due to [55] to multiple and intersectional groups. For this model, [19] show that for any set of utilities and amount of bias, there are utility-dependent non-intersectional constraints that achieve optimal latent utility for the ranking problem. However, since their constraints are a function of the latent utilities, which are not observed, these constraints cannot be determined in practice. While we also consider the model of implicit bias [19] introduced, the intersectional constraints we propose are different and do not depend on the, unknown, latent utilities. [35] study selection under a different model of bias, where the decision maker's observed utility has higher than average noise for individuals in the affected group. They consider a family of constraints and show that, for a single affected group, these constraints increase the latent utility. Unlike them, we consider multiple and intersectional groups and study a different model of bias. Finally, unlike these prior works, we also study the maximum utility achievable by using non-intersectional constraints.

*Intersectionality as a source of bias.* The discussion of being subjected to multiple biases has originally focused on the experience of Black women versus that of non-Black women and Black men. The "double jeopardy" and "multiple jeopardy" hypotheses posit that belonging to more than one affected group – as is the case for Black women – disproportionately increases the experienced bias [36, 54]. Empirically, for example, [34] show that returns to schooling in sub-Saharan Africa depend both gender and ethnicity and [84] find that peer-reviewed scores for post-doctoral fellowships were a function of both the candidate's gender and their affiliation with reviewers. Since [36, 54], several works have proposed extensions of this theory beyond two socially-salient attributes; e.g., [14, 23, 25–27, 69]. The implicit bias models that we consider in this work can be viewed as motivated by these works, and in particular, by the multiple jeopardy model [54].

*Intersectionality vs. non-intersectionality.* To the best of our knowledge, there are only few examples of mathematical studies that analyze how the belonging to intersectional groups interacts with policies. [3] study how individuals suppress or foster different dimensions of their identity to increase economic reward. [17] study the effect of intersectional vs. non-intersectional interventions on the share of different groups among the selected individuals over time. Our focus here is understanding the advantages of intersectional constraints over dimension-by-dimension non-intersectional constraints – albeit in the very different setup of selection under implicit bias.

# 2 MODEL

*Notation.* For a number $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. $\mathcal{U}$ denotes the uniform distribution over $[0, 1]$. We use $w \sim \mathcal{D}$ notation to denote that $w$ is an independent sample from distribution $\mathcal{D}$. For a distribution $\mathcal{D}$ over $\mathbb{R}$, we use $\mu_{\mathcal{D}} \colon \mathbb{R} \to \mathbb{R}_{\geq 0}$ to denote its probability density function and $F_{\mathcal{D}} \colon \mathbb{R} \to [0, 1]$ to denote its cumulative distribution function. We say a distribution $\mathcal{D}$ over $\mathbb{R}$ is continuous if $\mu_{\mathcal{D}}$ exists and is finite at all points in $\mathbb{R}$. The support of a continuous distribution $\mathcal{D}$ over $\mathbb{R}$ is the set $\{x \in \mathbb{R} \colon \mu_{\mathcal{D}}(x) > 0\}$, and is denoted by $\mathrm{supp}(\mathcal{D})$. Given two vectors $x, y \in \mathbb{R}^m$, we use $\langle x, y \rangle$ to denote their inner product $\sum_{i=1}^{m} x_i y_i$.

## 2.1 Selection Problem

The task of selecting a subset of individuals from a pool of applicants or employees arises in many contexts such as hiring, college admission, and selection for fellowships or board of directors. In these settings, the basic mathematical problem is as follows: Given a number $m$ and for each of the $m$ individuals, or more generally *items*, $i \in [m]$ a non-negative *latent utility* $w_i \geq 0$, the set selection problem asks to find a subset of $n$ items that has the maximum sum of latent utilities. If we represent a subset by a binary vector $x \in \{0, 1\}^m$, where $x_i = 1$ indicates that $i$ is in the subset and $x_i = 0$ indicates otherwise, the goal is to find $x \in \{0, 1\}^m$ such that $\langle x, w \rangle$ is maximized subject to $\sum_{i=1}^{m} x_i = n$.

## 2.2 Affected Groups, Intersections, and a Model of Implicit Bias

We consider the setting with $p$ affected groups (henceforth referred to as just groups) $G_1, G_2, \ldots, G_p \subseteq [m]$. Each of the $m$ items may belong to one or more of the $p$ groups that face implicit bias, or may belong to none of these groups, i.e., in $[m] \setminus \left(G_1 \cup G_2 \cup \cdots \cup G_p\right)$, and hence, not face any implicit bias. These groups can intersect arbitrarily. We use the following notation to capture all the intersections that can arise from $p$ groups: For a set $S \subseteq [p]$ of groups, let $\sigma \in \{0,1\}^p$ denote the corresponding indicator vector, i.e., for all $\ell \in [p]$, $\sigma_\ell = 1$ if $\ell \in S$ and $\sigma_\ell = 0$ otherwise. Let $I_\sigma \subseteq [m]$ denote the set of elements that belong to every group in $S$ and none of the groups not in $S$. Formally, when $S \neq \emptyset$ and, hence, $\sigma$ is not the all 0s vector (denoted by 0), let $I_\sigma := \left(\bigcap_{\ell:\sigma_\ell=1} G_\ell\right) \setminus \left(\bigcup_{\ell:\sigma_\ell=0} G_\ell\right)$. For $\sigma = 0$, let $I_0 := [m] \setminus \left(G_1 \cup G_2 \cup \cdots \cup G_p\right)$ denote the items in none of the $p$ groups. Where with some abuse of notation we used $I_0$ to denote $I_{00}$ for $p = 2$, $I_{000}$ for $p = 3$, and so on. Thus, the sets $\{I_\sigma\}_{\sigma \in \{0,1\}^p}$ partition the set of items $[m]$.

We focus on the extension of the implicit bias model of [55] presented in [19]. Later, in Section 3.3, we also consider further generalizations of this model. In this model, the decision maker does not observe the latent utilities of the items. Instead, for each item $i$, they see an observed utility $\widehat{w}_i$, which is their possibly biased estimate of $w_i$. In particular, the decision maker might perceive that items belonging to certain groups have a lower observed utility: $\widehat{w}_i < w_i$. For each group $\ell \in [p]$, there is an implicit bias parameter $0 < \beta_\ell \leq 1$ that captures the relative implicit bias faced by items in $G_\ell$ compared to items not in $G_\ell$. The total implicit bias experienced by an item $i$ is assumed to be the product of the implicit bias parameters of all groups it belongs to: $\prod_{\ell \in [p]:\, G_\ell \ni i} \beta_\ell$. Thus, for a given latent utility $w_i$ of item $i$, the observed utility is

$$\widehat{w}_i := \left(\prod\nolimits_{\ell \in [p]:\, G_\ell \ni i} \beta_\ell\right) \cdot w_i. \qquad \text{(Implicit bias)} \quad (1)$$

The property that individuals belonging to multiple groups are subject to more acute implicit bias is motivated by similar observations in the real world, which have been reproduced across many contexts such as hiring in industry, promotions in industry and academia, and peer-review in academia [30, 31, 72, 84] It also aligns with [54], which proposes a multiplicative-model where individuals face the compounded effect of the biases of groups they belong to.

As an illustration of this intersectional implicit bias model, consider two groups where the implicit bias parameter of the first group is $\beta_1$ and of the second group is $\beta_2$. An item which belongs to both $G_1$ and $G_2$ (i.e., in $I_{11}$) experiences an implicit bias $\beta_1\beta_2$. Whereas items in $G_1$ and not $G_2$ (i.e., in $I_{10}$) experience an implicit bias $\beta_1$ and items in $G_2$ but not $G_1$ (i.e., in $I_{01}$) experience an implicit bias $\beta_2$. Items neither $G_1$ nor $G_2$ (i.e., in $I_{00}$) do not face implicit bias. As a numerical example, if $\beta_1 = 0.9$ and $\beta_2 = 0.8$, then an item $i \in I_{11}$ experiences an implicit bias 0.72, which is more acute than the implicit bias experienced by items in $I_{10}$ or in $I_{10}$, which experience biases 0.9 and 0.8 respectively.

In Section 3, we consider a generalization of this model, where for each intersection $I_\sigma$, there is an increasing function $b_\sigma$, and the observed utility of an item $i$ in intersection $I_\sigma$ is $\widehat{w}_i = b_\sigma(w_i)$. This generalization captures Equation (1) when $b_\sigma(x) := x \cdot \prod_{\ell \in [p]:\, \sigma_\ell = 1} \beta_\ell$ for all intersections $\sigma$ and $x \geq 0$.

## 2.3 Selection Under Implicit Bias: Intersectional and Non-intersectional Constraints

The decision maker would like to pick an $x^\star \in \{0,1\}^m$ with $n$ items, i.e., $\sum_{i=1}^m x_i^\star = n$, that maximizes latent utility:

$$x^\star := \operatorname{argmax}_{x \in \{0,1\}^m} \langle x, w \rangle, \quad \text{s.t.,} \quad \sum_{i=1}^m x_i = n. \qquad (2)$$

However, due to their implicit bias, the decision maker instead maximizes the observed utility: They choose a selection $\widehat{x}$ with $n$ items that maximizes the observed utility,

$$\widehat{x} := \operatorname{argmax}_{x \in \{0,1\}^m} \langle x, \widehat{w} \rangle, \quad \text{s.t.,} \quad \sum_{i=1}^m x_i = n. \qquad (3)$$

Since $\widehat{x}$ maximizes a different objective than $x^\star$, it could be very different from $x^\star$. Hence, it may have a much smaller latent utility: $\langle \widehat{x}, w \rangle \ll \langle x^\star, w \rangle$. To see this, consider two items where the latent utility of the first item is $w_1 = 1$ and of the second item is $w_2 = 0.1$. Suppose the decision maker has to select one item (i.e., $n = 1$) and there are two groups. Since $x^\star$ maximizes the latent utility, it selects the first item. However, if the first item is in $I_{11}$ and the second item is in $I_{00}$, then their observed utilities are $\widehat{w}_1 = \beta_1\beta_2$ and $\widehat{w}_2 = 0.1$ respectively. If $\beta_1\beta_2 < 0.1$, then $\widehat{x}$ selects the second item. Hence, it has a latent utility $\langle \widehat{x}, w \rangle = 0.1$, which is significantly smaller than the latent utility of $x^\star$: $\langle x^\star, w \rangle = 1$. Apart from this, $\widehat{x}$ also selected fewer items facing implicit bias, i.e., items $i$ for which $\widehat{w}_i < w_i$; adversely affecting such items.

Affirmative action policies broadly seek to reduce and counter implicit bias and other systematic biases. There are several types of affirmative action policies [6, 78]. Here, we consider lower-bound constraints, such as the Rooney Rule, which require the decision maker to select at least a specified number of candidates from each group. Recent works [19, 35, 55] on decision-making in the presence of implicit bias have shown that in the setting with a single group, such constraints can serve an additional purpose: to improve the latent utility of the subset selected by the decision maker. Motivated by this, we study the efficacy of these constraints to improve the latent utility in the setting with multiple and intersectional groups, as discussed in Section 2.2. Specifically, we consider two types of lower-bound constraints: non-intersectional and intersectional:

- *Non-intersectional constraints* are specified by lower bounds $L_1, L_2, \ldots, L_p \in \mathbb{Z}_{\geq 0}$, and require that, for each $\ell \in [p]$, the decision maker include at least $L_\ell$ items from group $G_\ell$.
- *Intersectional constraints* are specified by $2^p$ lower bounds, $L_\sigma \in \mathbb{Z}_{\geq 0}$ for each $\sigma \in \{0,1\}^p$, and require that, for each $\sigma \in \{0,1\}^p$, the decision maker include at least $L_\sigma$ items from intersection $I_\sigma$.

When the context is clear, we use $L$ to denote the vector of lower bounds in either intersectional or non-intersectional constraints. Given a vector $L$ defining a lower-bound constraint, either non-intersectional or intersectional, let $C(L) \subseteq \{0,1\}^m$ be the set of all subsets with $n$ items that satisfy the constraints defined by $L$. Given a constraint $L$, the *constrained* decision maker chooses the selection $\widetilde{x}$ with the highest observed utility in $C(L)$:

$$\widetilde{x} := \operatorname{argmax}_{x \in C(L)} \langle x, \widehat{w} \rangle. \qquad (4)$$

Our goal is to understand how close to the optimal latent utility, $\langle x^\star, w\rangle$, can the decision maker (who selects) get when a policy maker (who decides the lower-bound constraints) imposes intersectional vs. non-intersectional constraints on them. In particular, given groups $G_1, \ldots, G_p$ and some unknown implicit bias parameters $\beta_1, \ldots, \beta_p$, which lower bound vectors $L$ have the property that the latent utility of the selection $\widetilde{x}$, $\langle \widetilde{x}, w\rangle$, is close to $\langle x^\star, w\rangle$.

## 2.4 Utility Ratio

As in [19, 35, 55], we study the setting where latent utilities of all items are drawn independently from some continuous distribution $\mathcal{D}$ with a non-negative support. We begin with the uniform distribution and later consider other distributions in Section 3.3. To measure the relative efficacy of different constraints, we consider a *utility ratio* defined as the expected value of the ratio of $\langle \widetilde{x}, w\rangle$ to $\langle x^\star, w\rangle$. Here, $x^\star$ is the selection with the highest latent utility (as defined in Equation (2)) and $\widetilde{x}$ is the selection picked by the constrained decision maker (as defined in Equation (4)). For a draw of latent utilities $w$, $\frac{\langle \widetilde{x}, w\rangle}{\langle x^\star, w\rangle}$ is the fraction of the optimal latent utility obtained by the constrained decision maker. Hence, the utility ratio measures the expected fraction of the optimal utility obtained by the constrained decision maker. Formally, given lower bounds $L$ and implicit bias parameters $\beta$, the utility ratio is:

$$\mathscr{R}_{\mathcal{D}}(L, \beta) \coloneqq \mathbb{E}_{w\sim\mathcal{D}}\left[\frac{\langle \widetilde{x}, w\rangle}{\langle x^\star, w\rangle}\right]. \qquad \text{(Utility ratio)} \quad (5)$$

Where $x^\star$ is a function of the utilities $w$ and $\widetilde{x}$ is a function of utilities $w$, implicit bias parameters $\beta$, and lower bounds $L$. It can be shown that the utility ratio is invariant to scaling of utilities (Proposition B.2), and its range is invariant across all distributions with the same mean and support (Proposition B.1). In particular, for the uniform distribution on $[0, C]$, for any $C > 0$, the range is $\frac{1}{2}$ to 1 (i.e., $\frac{1}{2} \le \mathscr{R}_{\mathcal{U}}(L, \beta) \le 1$). When discussing the uniform distribution on $[0, 1]$ we drop the subscript of $\mathscr{R}$ and use $\mathscr{R}(L, \beta)$ to denote $\mathscr{R}_{\mathcal{U}}(L, \beta)$.

# 3 RESULTS

## 3.1 Sub-Optimal Latent Utility for Any Non-intersectional Constraints

Our first result establishes an upper bound on the maximum utility ratio (Equation (5)) that a policy maker can secure in the presence of implicit bias by using non-intersectional constraints (Section 2.3). For simplicity, we first consider the case where utilities are distributed according to the uniform distribution on $[0, 1]$ and, later, consider generalizations to other distributions in Section 3.3. By definition, the utility ratio $\mathscr{R}(L, \beta)$ is at most 1. We show that for two groups (i.e., $p = 2$), if the utilities are independently drawn from the uniform distribution, then no matter which non-intersectional lower bounds $L_1 \ge 0$ and $L_2 \ge 0$ the policy maker chooses, $\mathscr{R}(L, \beta)$ is bounded away from 1 whenever $\beta_1 < 1$ and $\beta_2 < 1$.

Theorem 3.1 (**Non-intersectional constraints cannot recover full utility**). *Suppose the latent utilities are uniformly distributed on* $[0, 1]$*, the fraction of candidates selected is between* $\eta$ *and* $1 - \eta$ *for some constant* $\eta > 0$ *(i.e.,* $\eta < \frac{n}{m} < 1 - \eta$*), and the size of each intersection is greater than* $\rho m$ *(i.e., for all* $\sigma \in \{0, 1\}^2$*,* $|I_\sigma| > \rho m$*). For all implicit bias parameters* $0 < \beta_1, \beta_2 < 1$ *and constants* $\eta > 0$ *and* $\rho > 0$*, there is a threshold* $m_0 \in \mathbb{N}$ *such that if the number of candidates is more than this threshold,* $m \ge m_0$*, then for any non-intersectional lower bounds* $L_1, L_2 \ge 0$ *the utility ratio is strictly smaller than one, where the difference between* 1 *and the utility ratio depends up on* $\eta, \rho, \beta_1$*, and* $\beta_2$ *as follows:*

$$\mathscr{R}(L, \beta) \le 1 - \left(\frac{\rho}{3} \cdot \min\{\eta, 1 - \eta\} \cdot (1 - \beta_1) \cdot (1 - \beta_2)\right)^2. \quad (6)$$

Thus, Theorem 3.1 establishes that for uniformly distributed utilities, a policy maker cannot recover the full utility ratio with non-intersectional constraints for any bias parameters $0 < \beta_1, \beta_2 < 1$. As an example, suppose $\eta = \frac{1}{2}$ and $\rho = \frac{1}{4}$, then Equation (6) says that $\mathscr{R}(L, \beta) \le 1 - \left(\frac{1}{24}(1 - \beta_1) \cdot (1 - \beta_2)\right)^2$, which is strictly smaller than 1 for any $\beta_1 < 1$ and $\beta_2 < 1$ and is independent of the number of candidates $m$. This is in contrast to the setting with a single group, where non-intersectional constraints can recover a utility ratio arbitrarily close to 1 as the number of candidates $m$ increase [19, 35, 55].

We emphasize that the upper bound in Theorem 3.1 does not depend on the specific groups $G_1$ and $G_2$ or the specific value of $n$: it only requires that the intersection sizes are not too small and $n$ is not too close to 0 or $m$. For instance, in several admissions and hiring contexts, one can expect $\eta$ non-vanishing as even the most selective undergraduate programs in the US select more than 5% of the total applicants, which corresponds to $\eta \ge \frac{1}{20}$ [32]. Moreover, a majority of US undergraduate programs select between 0.4 and 0.8 fraction of the total applicants, which corresponds to $\frac{2}{5} \le \eta \le \frac{4}{5}$ [32].

Theorem 3.1 immediately implies the same result for $p > 2$, for instance, by adding empty groups or repeating the family of groups. In summary, Theorem 3.1 shows that non-intersectional constraints are not sufficient to recover the entire latent utility in the presence of multiple and intersectional groups. The dependence on implicit bias parameters $\beta$, and bounds on intersection size $\rho$ and selection rate $\eta$ are also natural and we discuss these below.

*Dependence on* $\beta$. One can verify that as $(\beta_1, \beta_2) \to (1, 1)$, the upper bound in Theorem 3.1 goes to 1. In particular, it becomes vacuous at $\beta_1 = \beta_2 = 1$. This is expected as when $\beta_1 = \beta_2 = 1$, no item faces implicit bias: for all items $i$, $w_i = \widehat{w}_i$. Further, the upper bound in Theorem 3.1 also goes to 1, as $\beta_1 \to 1$ while $\beta_2 \in (0, 1)$ is fixed (and vice versa). This holds as when $\beta_1 = 1$, only one group faces implicit bias. Thus, one can use the constraints given by [19], for the $p = 1$ case, to recover the near-optimal utility.

*Dependence on* $\rho$. The upper bound also goes to 1 as $\rho \to 0$. In particular, it becomes vacuous when $\rho = 0$. This is expected as when $\rho = 0$, $G_1$ and $G_2$ may not intersect. If the intersection is empty, then non-intersectional and intersectional constraints are the same, and one can use intersectional constraints in Theorem 3.2 to recover the near-optimal utility.

*Dependence on* $\eta$. The upper bound in Theorem 3.1 also goes to 1 as $\eta$ approaches either 0 or 1, while $\rho \in (0, 1)$ is fixed. This is because when $\eta = 1$, then both $\widetilde{x}$ and $x^\star$ select all $m$ items and, hence, $\widetilde{x} = x^\star$. Therefore, when $\eta = 1$, the utility ratio is 1. When $\eta$ is close to 0, $n$ is significantly smaller than $m$ and, hence, the best $n$ items in each intersection have latent utility very close to 1 with high probability. In this case, even if a decision maker is extremely biased, say they only select items from one intersection (e.g., White men), they would select $n$ items whose latent utility is very close to

$n$, which is the maximum value of the latent utility when utilities are drawn from the uniform distribution.

The basic property used to prove Theorem 3.1 is that: Any selection which deviates from proportional representation has a latent utility smaller than optimal. At a high level, Theorem 3.1 holds because the decision maker can alter their selection by selecting fewer (respectively more) items from $I_{11} \coloneqq G_1 \cap G_2$ and more (respectively fewer) items from $I_{01}$ and $I_{01}$, while keeping the number of selections from $G_1$ and from $G_2$ invariant. In the proof, we show that for any $L_1$ and $L_2$, the decision maker selects a significantly less-than-proportional number of candidates from at least one of the four intersections. The proof of Theorem 3.1 appears in Supplementary Material A.2 and an overview of the proof appears in Section 4.1.

## 3.2 Optimal Latent Utility With Intersectional Constraints

Complementing Theorem 3.1, we show that, if the policy maker is allowed to place constraints on intersections, then they can recover all but a vanishing (with $m$) fraction of the total latent utility. In particular, for any desired constant $0 < \varepsilon < 1$, number of items $m$, and groups $G_1, \ldots, G_p$, we give lower bounds $L_\sigma \geq 0$, for each intersection $\sigma$, such that for *any* implicit bias parameters $0 < \beta_1, \ldots, \beta_p \leq 1$, constrained selection leads to a utility ratio more than $1 - \varepsilon$.

Theorem 3.2 (**Intersectional constraints can (asymptotically) recover full utility**). *Suppose that the fraction of selected candidates is greater than $\eta$ for some constant $\eta > 0$ (i.e., $\frac{n}{m} > \eta \cdot m$). For all constants $0 < \varepsilon < 1$, $\eta > 0$, and number of groups $p \in \mathbb{N}$, there exists a threshold $m_0 \in \mathbb{N}$, such that if the number of candidates is more than this threshold, $m \geq m_0$, then for any groups $G_1, \ldots, G_p \subseteq [m]$, there exist intersectional constraints, $L_\sigma \geq 0$ for each intersection $\sigma \in \{0,1\}^p$, which for any implicit bias parameters $0 < \beta_1, \ldots, \beta_p \leq 1$ and any continuous distribution $\mathcal{D}$ with non-negative support, have a utility ratio at least $1 - \varepsilon$, i.e., $\mathscr{R}_{\mathcal{D}}(L, \beta) \geq 1 - \varepsilon$.*

Thus, Theorem 3.2 shows that a policy maker using intersectional constraints can recover utility ratio arbitrarily close to 1 for any bias parameters $0 < \beta_1, \beta_2 < 1$. In contrast to Theorem 3.1, here, the difference between the utility ratio and 1 approaches 0 as the number of candidates increase. Moreover, Theorem 3.2 holds for a larger choice of parameters than Theorem 3.1: (1) Any constant $\eta$ and sizes of the intersections ($\forall \sigma \in \{0,1\}^m$, $|I_\sigma|$) that satisfy the conditions in Theorem 3.1 also satisfy the conditions in Theorem 3.2, and (2) while $\mathcal{D}$ is fixed to be the uniform distribution on $[0,1]$ in Theorem 3.1, $\mathcal{D}$ can be the uniform distribution on $[0,1]$ or any other continuous distribution in Theorem 3.2.

We note that the intersectional constraints promised in Theorem 3.2 do not depend on the values of the implicit bias parameters $\beta_1, \ldots, \beta_p$ and the distribution of latent utility $\mathcal{D}$ and only depend on the intersection sizes, i.e., $|I_\sigma|$ for all $\sigma \in \{0,1\}^p$. Thus, they are applicable in practice when the amount of implicit bias and utility distributions are not know. Moreover, independence from $\beta_1, \ldots, \beta_p$ allows the same constraints to be used for different decision makers – who may have different implicit biases – and independence from $\mathcal{D}$ allows them to be used across selection tasks with different utility distributions. Finally, the independence also allows the constraints to be stable over time, while the implicit bias [21, 74] and utility distributions may change [33]. The proof of Theorem 3.2 appears in Supplementary Material A.3 and its overview appears in Section 4.2.

*Independence from $\beta$ and $\mathcal{D}$.* The independence from $\beta_1, \ldots, \beta_p$ relies on the fact that if the optimal constrained selection vector $\widetilde{x}$ picks $k \in \mathbb{N}$ items from an intersection, then these are the $k$ items with the highest latent utility in the intersection, irrespective of the value of $\beta_1, \ldots, \beta_p$. The independence from $\mathcal{D}$ relies on the fact that the sets of items selected by $x^\star$ and $\widetilde{x}$ from an intersection only depend on the order of the latent utility of items and not the actual values of latent utilities of items. Note, however, that while the lower bound on the utility ratio is independent of $\mathcal{D}$, the actual utility achieved by the non-intersectional constraints depends on $\mathcal{D}$.

*Computational Complexity.* Since each candidate belongs to exactly one intersection and there are $m$ candidates, there are at most $m$ non-empty intersections. Hence, even when the total number of intersections $2^p$ is larger than $m$, the decision maker has to consider at most $m$ (non-empty) intersections (as they cannot select candidates from empty intersections). Using this, we can show that the decision maker can find $\widetilde{x}$ in $O(m \log m)$ arithmetic operations.

## 3.3 Extensions of Theorem 3.1 and Theorem 3.2

*Extensions of Theorem 3.1.* A natural question raised by Theorem 3.1 is: How small can the upper bound on $\mathscr{R}(L, \beta)$ be? We show that $\mathscr{R}(L, \beta) \leq \frac{8}{9} + \frac{3}{2} \cdot \max(\beta_1, \beta_2)$. While this bound becomes vacuous when either $\beta_1 \geq \frac{2}{27}$ or $\beta_2 \geq \frac{2}{27}$, in the regime of large implicit bias, i.e., when $\beta_1$ and $\beta_2$ are close to 0, this bound approaches $\frac{8}{9}$.

Proposition 3.3. *Suppose the latent utilities are uniformly distributed on $[0,1]$, the fraction of candidates selected is $\frac{1}{2}$ (i.e., $\frac{n}{m} = \frac{1}{2}$), and $G_1$ and $G_2$ are such that all intersections have size $\frac{m}{4}$ (i.e., for all $\sigma \in \{0,1\}^2$, $|I_\sigma| = \frac{m}{4}$). For all implicit bias parameters $0 < \beta_1, \beta_2 \leq 1$ and for all non-intersectional lower-bound constraint $L_1, L_2 \geq 0$, there exists an threshold $m_0 \in \mathbb{Z}$, such that for all $m \geq m_0$, the utility ratio is upper bounded by a quantity that approaches $\frac{8}{9}$ as $\beta_1$ and $\beta_2$ go to 0, where the specific dependence of the upper bound on $\beta_1$ and $\beta_2$ is as follows: $\mathscr{R}(L, \beta) \leq \frac{8}{9} + \frac{3}{2} \cdot \max(\beta_1, \beta_2)$.*

We do not know if the bound in Proposition 3.3 is tight but conjecture that it cannot go below $\frac{8}{9}$ for the uniform distribution. The proof of Proposition 3.3 appears in Supplementary Material A.4.

Another question is if Theorem 3.1 is specific to the uniform distribution of utilities or to the implicit bias model in Equation (1). Other relevant distribution families include power-law and truncated normal. These families arise in peer-review and university admissions as distributions of citations [22] and test scores, e.g., SAT [33], respectively. As for the implicit bias model, Equation (1) assumes that implicit bias experienced by an item $i$ in two groups, say, $G_1$ and $G_2$ is exactly the product of the implicit bias parameters $\beta_1$ and $\beta_2$ of these groups, i.e., $\beta_1\beta_2$. But depending on the specific context, the implicit bias could be more severe, e.g., $(\beta_1\beta_2)^2$, or less severe, e.g., $(\beta_1\beta_2)^{\frac{1}{2}}$.

We show that a version of Theorem 3.1 holds for truncated power-law and normal distributions and the above implicit bias models. We consider a family of models where the implicit bias at each intersection $\sigma$ is given by a strictly increasing function

$b_\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, such that, the observed utility of each item $i \in I_\sigma$ in this intersection is:

$$\widehat{w}_i \coloneqq b_\sigma(w_i). \tag{7}$$

This model captures the model in Equation (1) when $b_\sigma(x) \coloneqq x \cdot \prod_{\ell \in [p]:\ \sigma_\ell = 1} \beta_\ell$. Further, when $p = 2$, changing $b_{11}(x) \coloneqq x \cdot (\beta_1\beta_2)^2$ or $b_{11}(x) \coloneqq x \cdot (\beta_1\beta_2)^{\frac{1}{2}}$ in the previous sentence captures the models mentioned above. We prove a version of Theorem 3.1 under the following assumption:

Assumption 1. *There are constants $c, d > 0$ such that the probability density function of $\mathcal{D}$, $\mu_{\mathcal{D}}$, is differentiable and for all $x$ in the support of $\mathcal{D}$ and takes values between $c$ and $\frac{1}{c}$ (i.e., $c \leq \mu_{\mathcal{D}}(x) \leq \frac{1}{c}$), and the bias function for each intersection (i.e., $b_\sigma$ for each $\sigma \in \{0,1\}^p$) is differentiable, takes value at most $\frac{1}{c}$ (i.e., $b_\sigma(x) \leq \frac{1}{c}$), and has a derivative between $d$ and $\frac{1}{c}$ (i.e., $d \leq b'_\sigma(x) \leq \frac{1}{c}$) for each $x$ in the support of $\mathcal{D}$.*

The assumption on $b'_\sigma$ ensures that two items with similar latent utilities also have similar observed utilities: if $|w_i - w_j| = \varepsilon$ then $d\varepsilon \leq |\widehat{w}_i - \widehat{w}_j| \leq \frac{\varepsilon}{c}$. The functions $b_\sigma$ show up in the generalization of Theorem 3.1, and the upper bound on $b_\sigma$ is needed to re-scale the right-hand side of Equation (8) to be non-negative. The assumption on $\mathcal{D}$ is satisfied, for instance, by the truncated normal and power-law distributions: This is because their probability density functions are bounded between two positive constants at any point in the support. As opposed to them, the *un-truncated* normal and power-law distributions do not satisfy Assumption 1, this is because their probability density function, $\mu$, approaches 0 for large values (i.e., $\mu(x) \to 0$ as $x \to \infty$) and, hence, does not satisfy the lower bound. Because of this, for these distributions, the constant $c$ decreases as the upper-end of the truncation interval increases. Further, the extensions of the model in Equation (1) discussed above also satisfy Assumption 1: In particular, if $\mathcal{D}$ is the uniform distribution on $[0,1]$ then Assumption 1 holds for $c = 1$ and for $d = (\beta_1\beta_2)^2$ (and $d = (\beta_1\beta_2)^{\frac{1}{2}}$ respectively).

Theorem 3.4. *Suppose the fraction of candidates selected is between $\eta$ and $1-\eta$ for some constant $\eta > 0$ (i.e., $\eta < \frac{n}{m} < 1-\eta$), and the size of each intersection is at least $\rho m$ (i.e., for all $\sigma \in \{0,1\}^2$, $|I_\sigma| > \rho m$). Under the implicit bias model in Equation (7), for any continuous distribution $\mathcal{D}$ with non-negative support and set of strictly increasing functions $b_\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, where there is one function for each intersection $\sigma \in \{0,1\}^2$, if Assumption 1 holds, then there is a threshold $m_0 \in \mathbb{N}$ such that when the number of candidates is more than this threshold, $m \geq m_0$, any non-intersectional lower-bound constraint $L_1, L_2 \geq 0$ the utility ratio $\mathscr{R}_{\mathcal{D}}(L,\beta)$ is at-most*

$$1 - \left(c^4 \rho \min\{\eta, 1-\eta\} (b_{00} - b_{10} - b_{01} + b_{11}) \circ \left(F_{\mathcal{D}}^{-1}\left(1 - \frac{n}{m}\right)\right)\right)^2.^1 \tag{8}$$

Thus, for any bias functions and distributions for which $(b_{00} - b_{10} - b_{01} + b_{11}) \circ \left(F_{\mathcal{D}}^{-1}\left(1 - \frac{n}{m}\right)\right)$ is non-zero, Theorem 3.4 shows that the utility ratio is strictly smaller than 1 for any non-intersectional lower bounds $L_1$ and $L_2$. Complementing this, we show that when this specific additive quantity is 0, there are non-intersectional constraints which recover near-optimal latent utility (Proposition A.27).

[1] Given functions $f_1$, $f_2$, $f_3$, and $f_4$, and a number $x \in \mathbb{R}$, we use $(f_1 + f_2 + f_3 + f_4) \circ (x)$ to denote $f_1(x) + f_2(x) + f_3(x) + f_4(x)$.

We present the proof of Theorem 3.4 in Supplementary Material A.5 and discuss how it differs from the proof of Theorem 3.1 in Section 4.1.

*Extensions of Theorem 3.2.* The proof of Theorem 3.2 only needs the following property of the implicit bias model: For two items in the same intersection, the order of their latent utilities is the same as the order of their observed utilities. Abstracting this leads us to the generalization of Theorem 3.2 to the implicit bias in Equation (7), which was discussed in Section 3.1. This, in particular, shows that intersectional constraints can have an arbitrarily high utility ratio for the generalization of the model of implicit bias considered in Theorem 3.4.

Corollary 3.5. *Suppose that the fraction of candidates selected is equal to some constant $\eta > 0$ (i.e., $\frac{n}{m} = \eta$). Given $0 < \varepsilon < 1$ and $\eta > 0$, consider the threshold $m_0$ in Theorem 3.2, for all $m \geq m_0$ and group structures $G_1, \ldots, G_p \subseteq [m]$, the intersectional lower bound $L$ from Theorem 3.2 are such that, for any set of strictly increasing functions $b_\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, where there is one function for each intersection $\sigma \in \{0,1\}^p$, they satisfy $\mathscr{R}_{\mathcal{D}}(L,\beta) \geq 1 - \varepsilon$.*

Finally, we state a desirable fairness property of the constraints in Theorem 3.2. For each intersection $\sigma \in \{0,1\}^p$, the constraints in Theorem 3.2 satisfy $L_\sigma \geq (1-\varepsilon) \cdot n \cdot \frac{|I_\sigma|}{m}$. Hence, for small $\varepsilon > 0$, the constrained decision maker must select at least a near-proportional number of items from each intersection. Thus, these constraints can be seen as a form of affirmative action.

# 4 OVERVIEW OF PROOFS

## 4.1 Proof Overview of Theorem 3.1

In this section, we explain the key ideas in the proof of Theorem 3.1 under some simplifying assumptions. Recall that utility ratio is $\mathscr{R}(L,\beta) \coloneqq \mathbb{E}_w\left[\frac{\langle \widetilde{x}, w\rangle}{\langle x^\star, w\rangle}\right]$, where $\widetilde{x}$ is the selection picked by the constrained decision maker (as defined in Equation (4)) and $x^\star$ is the selection with the highest latent utility (as defined in Equation (2)). We are given the implicit bias parameters $0 < \beta_1, \beta_2 \leq 1$ and the guarantee that for some $\eta > 0$ and $\rho > 0$,

$$\forall \sigma \in \{0,1\}^2, \quad \rho < \frac{|I_\sigma|}{m} < 1 - \rho \quad \text{and} \quad \eta < \frac{n}{m} < 1 - \eta. \tag{9}$$

For $\eta, \rho > 0$, let $\phi(\beta_1, \beta_2) \coloneqq \left(\frac{\rho}{3} \min\{\eta, 1-\eta\} (1-\beta_1)(1-\beta_2)\right)^2$. Our goal is to show that

$$\max_{L_1, L_2 \in \mathbb{Z}_{\geq 0}} \mathscr{R}(L,\beta) \leq 1 - \phi(\beta). \tag{10}$$

In this overview, our first simplifying assumption is that for all $L$ and $\beta$: $\mathscr{R}(L,\beta) \coloneqq \mathbb{E}_w\left[\frac{\langle \widetilde{x}, w\rangle}{\langle x^\star, w\rangle}\right] = \frac{\mathbb{E}_w[\langle \widetilde{x}, w\rangle]}{\mathbb{E}_w[\langle x^\star, w\rangle]}$. This assumption is *not true*. In the proof, we remove it by showing that the distribution of $\langle x^\star, w\rangle$ is concentrated around its mean. Since $x^\star$ is not a function of $L$, under the above assumption, it suffices to show that

$$\max_{L_1, L_2 \in \mathbb{Z}_{\geq 0}} \mathbb{E}_w\left[\langle \widetilde{x}, w\rangle\right] \leq (1 - \phi(\beta)) \cdot \mathbb{E}_w\left[\langle x^\star, w\rangle\right]. \tag{11}$$

**Challenges in computing $\mathbb{E}_w[\langle \widetilde{x}, w\rangle]$.** One approach to prove Equation (11) could be to compute $\mathbb{E}_w[\langle \widetilde{x}, w\rangle]$ explicitly as a function of $L$ and then verify Equation (11). This is the approach that [19] take for a single group. In their case, there is a simple iterative algorithm that, given $L_1$ and observed utilities, computes $\widetilde{x}$. The algorithm is as follows: Pick $L_1$ items with the highest observed utility from $G_1$ and from the remaining items pick $n - L_1$ items that

have the highest observed utility. [19] analyze this algorithm to compute $\mathbb{E}_w[\langle \widetilde{x}, w\rangle]$ as a function of $L_1$.

This algorithm and analysis straightforwardly extends to multiple *non-overlapping* groups: For each $\ell \in [p]$, pick $L_\ell$ items with the highest observed utility from $G_\ell$, then from the remaining items pick $n - L_1 - L_2 - \cdots - L_p$ items that have the highest observed utility. However, this algorithm breaks down when groups overlap. This is because items at intersections of multiple groups can satisfy multiple lower bounds. Moreover, at least in the case where the number of groups, $p$, is non-constant we do not expect there to be a simple algorithm which computes $\widetilde{x}$: This is because the NP-complete hitting set problem reduces to checking if $\widetilde{x}$ exists, i.e., if the specified lower bounds are satisfiable.

Instead of computing $\mathbb{E}_w[\langle \widetilde{x}, w\rangle]$ as a function of $L$, we express $\mathbb{E}_w[\langle \widetilde{x}, w\rangle]$ and $\mathbb{E}_w[\langle x^\star, w\rangle]$ as solution to two optimization programs. Then, we directly upper bound the ratio $\frac{\mathbb{E}_w[\langle \widetilde{x}, w\rangle]}{\mathbb{E}_w[\langle x^\star, w\rangle]}$ by analyzing the optimization programs.

*Step 1: Reduce computing $x^\star$ and $\widetilde{x}$ to a small number of variables.* A property of both $x^\star$ and $\widetilde{x}$ is that if they pick $k$ items from $I_\sigma$, then these are the $k$ items with the highest latent utility, or *equivalently* the highest observed utility, in $I_\sigma$ (see Observation A.6). Hence, determining $x^\star$ and $\widetilde{x}$ reduces to computing, the following quantities for each $\sigma \in \{0,1\}^2$ $K^\star_\sigma := \sum_{i \in I_\sigma} x^\star_i$ and $\widetilde{K}_\sigma := \sum_{i \in I_\sigma} \widetilde{x}_i$.

*Step 2: Express $K^\star$ and $\widetilde{K}$ as solutions to different optimization problems.* Since $x^\star$ and $\widetilde{x}$ are functions of randomly generated utilities, they and, hence, $K^\star$ and $\widetilde{K}$ are random variables. Under Assumption (9), we show that $K^\star$ and $\widetilde{K}$ are concentrated around the optimizers of optimization problems (13) and (14) respectively; where for any $\gamma \in [0,1]^4$ and $k \in \mathbb{R}^4_{\geq 0}$

$$f_\gamma(k) := \sum_\sigma \gamma_\sigma |I_\sigma| \cdot \int_{1-\frac{k_\sigma}{|I_\sigma|}}^{1} x dx. \tag{12}$$

$$\begin{aligned} \operatorname{argmax}_k \quad & f_1(k), \\ \text{s.t.,} \quad & \sum\nolimits_\sigma k_\sigma = n, \\ & \forall \sigma \in \{0,1\}^2, \quad 0 \leq k_\sigma \leq |I_\sigma|. \end{aligned} \tag{13}$$

$$\begin{aligned} \operatorname{argmax}_k \quad & f_\beta(k), \\ \text{s.t.,} \quad & k_{10} + k_{11} \geq L_1 \text{ and } k_{01} + k_{11} \geq L_2, \\ & \sum\nolimits_\sigma k_\sigma = n, \\ & \forall \sigma \in \{0,1\}^2,\ 0 \leq k_\sigma \leq |I_\sigma|. \end{aligned} \tag{14, 15}$$

While the integral in Equation (12) can be computed exactly, we use the integral-form because it generalizes to other distributions and bias functions, where the resulting integral may not be possible to compute. Let $x(k)$ be the selection that picks $k_\sigma$ items with the highest latent utility from $I_\sigma$ for all $\sigma \in \{0,1\}^2$. Suppose $k$ is feasible for Program (13) and Program (14). The constraints in Program (13) ensure $x(k)$ selects a total of $n$ candidates, and the additional constraints in Program (14) ensure $x(k)$ picks at least $L_1$ and $L_2$ candidates from $G_1$ and $G_2$. $f_\beta(k)$ roughly measures the expected *observed* utility of $x(k)$ and $f_1(k)$ roughly measures the expected *latent* utility $x(k)$:

$$f_\beta(k) = \mathbb{E}[\langle x(k), \widehat{w}\rangle] \pm O(m^{-1}) \;\&\; f_1(k) = \mathbb{E}[\langle x(k), w\rangle] \pm O(m^{-1}). \tag{16}$$

$f_1$ and $f_\beta$ can be shown to be strongly concave for all $\gamma$, and hence, Programs (13) and (14) have unique solutions. Formally, we prove the following concentration bound on $K^\star$ and $\widetilde{K}$.

Lemma 4.1. *Let $s^\star$ and $\widetilde{s}$ be the optimizers of Programs* (13) *and* (14) *respectively. With probability at least $1 - O(m^{-\frac{1}{4}})$,*

$$\forall \sigma \in \{0,1\}^2, \quad |K^\star_\sigma - s^\star_\sigma| \leq O(nm^{-\frac{1}{4}}) \;\&\; |\widetilde{K}_\sigma - \widetilde{s}_\sigma| \leq O(nm^{-\frac{1}{4}}).$$

Suppose that $K^\star$ and $\widetilde{K}$ are equal to minimizers of Programs (13) and (14) with probability 1. Then, from Equation (16),

$$\mathbb{E}\left[\langle x^\star, w\rangle\right] = f_1(K^\star) \pm O(m^{-1}) \;\&\; \mathbb{E}[\langle \widetilde{x}, w\rangle] = f_1(\widetilde{K}) \pm O(m^{-1}).$$

Suppose these hold with equality. Then because the feasible region of Program (13) is a superset of the feasible region of Program (14) and $K^\star$ maximizes $f_1$ over the feasible region of Program (13), it follows that $f_1(\widetilde{K}) = \mathbb{E}[\langle \widetilde{x}, w\rangle] \leq \mathbb{E}\left[\langle x^\star, w\rangle\right] = f_1(K^\star)$. The question is: Is $f_1(\hat{K})$ significantly smaller than $f_1(K^\star)$? Does Equation (17) (below) hold?

$$f_1(\widetilde{K}) \leq (1 - \phi(\beta)) \cdot f_1(K^\star). \tag{17}$$

*Step 3: Prove Equation* (17) *(Main argument).* Our first observation is that both $f_1$ is $\frac{1}{(1-\rho)m}$-strongly concave and $f_\beta$ is $\frac{1}{\rho m}$-Lipschitz continuous. Using the gradient test, we can analytically solve Program (13) to get $K^\star := \left\{|I_\sigma| \cdot \frac{n}{m}\right\}_{\sigma \in \{0,1\}^2}$. The claim follows because if $\widetilde{K}$ satisfies any inequality in Equation (15) with equality, then $\|K^\star - \widetilde{K}\|_2$ is large. Namely,

$$\|K^\star - \widetilde{K}\|_2^2 \geq 2nm(1-\rho) \cdot \phi(\beta). \tag{18}$$

Hence, by the $\frac{1}{(1-\rho)m}$-strong convexity of $f_1$ and the fact that $f(K^\star) \leq n$, Equation (17) holds. Otherwise if $\widetilde{K}$ satisfies all inequalities in Equation (15) with strict inequality, then because Program (14) has only three other constraints, while $\widetilde{K}$ has four coordinates, it follows that there is some constant $t_0 > 0$ and a vector, namely $v := (1,-1,-1,1)$ such that for all $-t_0 \leq t \leq t_0$, $\widetilde{K} + tv$ is feasible for Program (14). Since $\widetilde{K}$ is the optimal solution of Program (14), this implies that

$$\langle \nabla f_\beta(\widetilde{K}), v\rangle = 0. \tag{19}$$

Using the value of $K^\star$, we have

$$\left\langle \nabla f_\beta(K^\star), v\right\rangle = \left(1 - \frac{n}{m}\right) \cdot (1-\beta_1) \cdot (1-\beta_2). \tag{20}$$

This is sufficient to show that, in this case, $\|\nabla f_\beta(\widetilde{K}) - \nabla f_\beta(K^\star)\|_2$ is large:

$$\begin{aligned} \|\nabla f_\beta(\widetilde{K}) - \nabla f_\beta(K^\star)\|_2 &\geq \frac{1}{\|v\|_2} \cdot \left|\left\langle \nabla f_\beta(\widetilde{K}) - \nabla f_\beta(K^\star), v\right\rangle\right| \\ &\overset{(19),(20)}{=} \frac{1}{2}\left(1 - \frac{n}{m}\right) \cdot (1-\beta_1) \cdot (1-\beta_2). \end{aligned}$$

Combined with the fact that $f_\beta$ is $\frac{1}{\rho m}$-Lipschitz continuous, this implies that $\|K^\star - \widetilde{K}\|_2^2 \geq \frac{\rho^2 m^2}{4} \cdot \left(1 - \frac{n}{m}\right)^2 \cdot (1-\beta_1)^2 \cdot (1-\beta_2)^2 \geq 2mn(1-\rho) \cdot \phi(\beta)$. Thus, in this case also Equation (18) follows from $\frac{1}{(1-\rho)m}$-strong convexity of $f_1$.

*Generalization to other bias functions and distributions.* The proof of Theorem 3.4 is analogous to that of Theorem 3.1. Let $F_{\mathcal{D}}$ be the cumulative distribution function of $\mathcal{D}$ and $\mu_{\mathcal{D}}$ be the probability

density function of $\mathcal{D}$. The main difference is that $f_1$ and $f_\beta$ change to

$$f_1(k) \coloneqq \sum_\sigma |I_\sigma| \int_{z_\sigma(k)}^{z_\sigma(0)} x d\mu_{\mathcal{D}}(x), \quad \text{and}$$
$$f_b(k) \coloneqq \sum_\sigma |I_\sigma| \int_{z_\sigma(k)}^{z_\sigma(0)} b_\sigma(x) d\mu_{\mathcal{D}}(x) \tag{21}$$

where $z_\sigma(k) \coloneqq F_{\mathcal{D}}^{-1}\left(1 - \frac{k_\sigma}{|I_\sigma|}\right)$. We choose this definition of $f$ because of a similar reason: Under the general bias model and distribution, $f_b(k)$, roughly, measures the expected *observed* utility of $x(k)$ and $f_1(k)$, roughly, measures the expected *latent* utility $x(k)$. Next, we prove Lemma 4.1 for the new definition of $f$. The rest of the proof follows analogously once we prove that $f_1$ is $\frac{\Omega(1)}{(1-\rho)m}$-strongly concave and $f_b$ is $\frac{O(1)}{\rho m}$-Lipschitz continuous

## 4.2 Proof Overview of Theorem 3.2

In the proof we show that with high probability, for each intersection $\sigma \in \{0,1\}^p$: $\frac{\sum_{i\in I_\sigma} \widetilde{x}_i w_i}{\sum_{i\in I_\sigma} x_i^\star w_i} > 1 - \frac{\varepsilon}{2}$. This implies:

$$\frac{\langle \widetilde{x}, w\rangle}{\langle x^\star, w\rangle} = \frac{\sum_\sigma \sum_{i\in I_\sigma} \widetilde{x}_i w_i}{\sum_\sigma \sum_{i\in I_\sigma} x_i^\star w_i} > \frac{\sum_\sigma \sum_{i\in I_\sigma} x_i^\star w_i}{\sum_\sigma \sum_{i\in I_\sigma} x_i^\star w_i}\left(1 - \frac{\varepsilon}{2}\right) > 1 - \frac{\varepsilon}{2}. \tag{22}$$

The claimed result follows by taking the expectation of the above quantity. At a high level, our strategy is to find constraints such that $\widetilde{x}$ selects a similar number of items from each intersection as $x^\star$. This suffices to prove the result due to a property of the implicit bias model: If $\widetilde{x}$ picks $K$ items from any intersection $I_\sigma$, then these are the $K$ items with the highest latent utility in $I_\sigma$ (Observation A.6).

To see why this suffices, let $v_1 \le v_2 \le \cdots \le v_{|I_\sigma|}$ be the latent utilities of items in $I_\sigma$ in non-increasing order. Let $N$ and $\widetilde{N}$ be the random variables counting the number of items $x^\star$ and $\widetilde{x}$ select from $I_\sigma$. Using the above observation, we know that $x^\star$ and $\widetilde{x}$ select the items $v_1, v_2, \ldots, v_N$ and $v_1, v_2, \ldots, v_{\widetilde{N}}$ respectively. Hence,

$$\frac{\sum_{i\in I_\sigma} \widetilde{x}_i w_i}{\sum_{i\in I_\sigma} x_i^\star w_i} \ge \frac{\sum_{j=1}^{\widetilde{N}} v_j}{\sum_{i\in I_\sigma} x_i^\star w_i} = \frac{\sum_{j=1}^{\widetilde{N}} v_j}{\sum_{j=1}^{N} v_j} = \frac{\sum_{j=1}^{\widetilde{N}} v_j}{\sum_{j=1}^{\widetilde{N}} v_j + \sum_{j=\widetilde{N}+1}^{N} v_j}.$$

Using that for all $c \ge 0$, $\frac{x}{c+x}$ is an increasing function of $x$, we have

$$\frac{\sum_{i\in I_\sigma} \widetilde{x}_i w_i}{\sum_{i\in I_\sigma} x_i^\star w_i} \ge \frac{\widetilde{N} \cdot v_{\widetilde{N}}}{\widetilde{N} \cdot v_{\widetilde{N}} + \sum_{j=\widetilde{N}+1}^{N} v_j} \ge \frac{\widetilde{N} \cdot v_{\widetilde{N}}}{\widetilde{N} \cdot v_{\widetilde{N}} + (N - \widetilde{N}) \cdot v_{\widetilde{N}}} = \frac{\widetilde{N}}{N}. \tag{23}$$

Thus, if $\frac{\widetilde{N}}{N} > 1 - \frac{\varepsilon}{2}$ with high probability for all $\sigma \in \{0,1\}^p$, then for all $\sigma \in \{0,1\}^2$, it holds that $\frac{\sum_{i\in I_\sigma} \widetilde{x}_i w_i}{\sum_{i\in I_\sigma} x_i^\star w_i} > 1 - \frac{\varepsilon}{2}$. Towards this, we prove that for all $\sigma \in \{0,1\}^p$, $N_\sigma$ is concentrated around $|I_\sigma| \cdot \frac{n}{m}$:

Lemma 4.2. *For any fixed $\sigma \in \{0,1\}^p$ and $\Delta \ge 2$, it holds that $\mathbb{E}[N_\sigma] = |I_\sigma| \cdot \frac{n}{m}$ and $\Pr_w[N_\sigma > \mathbb{E}[N_\sigma] + \Delta] \le e^{-\frac{\Delta^2}{n}}$.*

Given Lemma 4.2, an obvious strategy is to set $L_\sigma = |I_\sigma| \cdot \frac{n}{m}$ for all $\sigma \in \{0,1\}^p$. However, this does not work because if $|I_\sigma|$ is small, then the concentration bound in Lemma 4.2 is weak. We overcome this by setting slightly larger bounds for small intersections and slightly smaller bounds for big intersections:

$$\text{for all } \sigma \in \{0,1\}^p, \qquad L_\sigma \coloneqq \frac{|I_\sigma|}{m} \cdot n \cdot (1-\varepsilon) + 2^{-p} n\varepsilon, \tag{24}$$

where if RHS is larger than $|I_\sigma|$, then we set $L_\sigma = |I_\sigma|$. Using Lemma 4.2, one can show that these constraints suffice.

# 5 CONCLUSIONS, LIMITATIONS, AND FUTURE WORK

In this work we studied the set selection problem in the presence of implicit bias where each item may belong to any intersectional socially-salient groups. Focusing on lower-bound constraints and randomly (and independently) drawn utilities, we first showed that no non-intersectional constraints achieve near-optimal utility. We then presented intersectional constraints that recover almost all utility for arbitrarily chosen biases for each intersectional group. Thus, the advantage of intersectional constraints is substantial over a dimension-by-dimension approach.

Our work raises several questions. While we show that the upper bound on the expected utility ratio in the non-intersectional case is 8/9 when $\beta$'s tend to zero and utilities have the uniform distribution, we do not know if this 8/9 is tight. Moreover, it would be interesting to investigate if this 8/9 bound can go further down for the distribution of utility changes or as the number of groups $p$ increases. Absent constraints, we know from Proposition B.1 that a ratio of $1/2$ can always be guaranteed for the uniform distribution.

More generally, our analysis suffers from several limitations, also sometimes found in closely related literature. First, the assumption of independent random draws of utilities is critical but may not be realistic in practice. Second, bias is assumed to be exogenous and due to the lack of a temporal element in bias, interventions cannot be evaluated in terms of their long-term equilibrium effects. Third, and relating to the previous point, the interaction between affirmative action via setting lower-bound constraints and the behavior of individuals or groups is not taken into account. To this end, the recently dropped law suit by the U.S. Department of Justice against Yale University[2], that accused admission to be biased against Asian-American and white American, exemplifies how by prioritizing some groups others necessarily feel or are deprioritized.

Moreover, lower-bound constraints – or affirmative action – is just one intervention to curb the adverse effects of implicit bias. Any effective approach against implicit bias must consider a diverse set interventions, including information campaigns and (re-)structured evaluations, complemented by increased accountability and transparency to enforce and guide these interventions. Here, our work also raises the question of how policy makers can incorporate intersectionality in other interventions and the potential advantages of doing so.

The presence of bias and discrimination unquestionably remains one of the pressing issues of our society. Implicit bias – vis-à-vis explicit discrimination – is too often viewed as an inevitable byproduct of human nature. To holistically address how policies can help rather than hinder the persistence of unjust outcomes it is important to unpack many connected questions: Which groups merit protection and when? How does affirmative action regarding one

[2] https://www.nytimes.com/2021/02/03/us/yale-admissions-affirmative-action.html

group affect another? Should interventions aim at reducing the existence of implicit bias or is it sufficient if they counter its effects?

## ACKNOWLEDGMENTS

This project is supported in part by an NSF Award (CCF-2112665). We would like to thank Jean-Paul Carvalho, Elisa Celis, and Patrick Loiseau for useful discussions.

## 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] Pragya Agarwal. 2020. *Sway: Unravelling Unconscious Bias.* Bloomsbury Publishing. https://books.google.com/books?id=Tt3LDwAAQBAJ
[3] Robert Akerlof. 2017. Value Formation: The Role of Esteem. *Games and Economic Behavior* 102 (2017), 1–19.
[4] Gordon W. Allport. 1954. *The Nature of Prejudice.* Addison-Wesley.
[5] Katherine Baldiga. 2014. Gender Differences in Willingness to Guess. *Management Science* 60, 2 (2014), 434–448.
[6] Surender Baswana, Partha Pratim Chakrabarti, Sharat Chandran, Yashodhan Kanoria, and Utkarsh Patange. 2019. Centralized Admissions for Engineering Colleges in India. *INFORMS Journal on Applied Analytics* 49, 5 (2019), 338–354.
[7] 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.
[8] 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.
[9] Scott Bland. 2017. Schumer to Introduce Rules for Diverse Senate Hiring. *Politico* (2017). https://www.politico.com/story/2017/02/schumer-diversity-nfl-rooney-rule-235477.
[10] Zachary Bleemer and Aashish Mehta. 2021. College Major Restrictions and Student Stratification. *UC Berkeley: Center for Studies in Higher Education* (2021).
[11] Avrim Blum and Kevin Stangl. 2020. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?. In *FORC (LIPIcs, Vol. 156).* Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 3:1–3:20.
[12] Tristan L. Botelho and Mabel Abraham. 2017. Pursuing Quality: How Search Costs and Uncertainty Magnify Gender-based Double Standards in a Multistage Evaluation Process. *Administrative Science Quarterly* 62, 4 (2017), 698–730. https://doi.org/10.1177/0001839217694358 arXiv:https://doi.org/10.1177/0001839217694358
[13] Tristan L Botelho and Marina Gertsberg. 2021. The Disciplining Effect of Status: Evaluator Status Awards and Observed Gender Bias in Evaluations. *Management Science* (2021).
[14] Rose M. Brewer, Cecilia A. Conrad, and Mary C. King. 2002. The Complexities and Potential of Theorizing Gender, Caste, Race, and Class. *Feminist Economics* 8, 2 (2002), 3–17.
[15] Irene Browne and Joya Misra. 2003. The Intersection of Gender and Race in the Labor Market. *Annual Review of Sociology* 29, 1 (2003), 487–513. https://doi.org/10.1146/annurev.soc.29.010202.100016 arXiv:https://doi.org/10.1146/annurev.soc.29.010202.100016
[16] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *FAT (Proceedings of Machine Learning Research, Vol. 81).* PMLR, 77–91.
[17] Jean-Paul Carvalho and Bary S. R. Pradelski. 2019. Identity and Underrepresentation: Interactions between Race and Gender. *Working paper* (2019).
[18] Marilyn Cavicchia. 2017. How to Fight Implicit Bias? With Conscious Thought, Diversity Expert Tells NABE. *American Bar Association* (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/
[19] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2020. Interventions for Ranking in the Presence of Implicit Bias. In *FAT**. ACM, 369–380.
[20] Elizabeth N. Chapman, Anna Kaatz, and Molly Carnes. 2013. Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities. *Journal of General Internal Medicine* 28, 11 (2013), 1504–10.
[21] Tessa E. S. Charlesworth and Mahzarin R. Banaji. 2019. Patterns of Implicit and Explicit Attitudes: I. Long-Term Change and Stability From 2007 to 2016. *Psychological Science* 30, 2 (2019), 174–192. https://doi.org/10.1177/0956797618813087 arXiv:https://doi.org/10.1177/0956797618813087 PMID: 30605364.
[22] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM review* 51, 4 (2009), 661–703.
[23] Patricia Hill Collins. 2000. Gender, Black Feminism, and Black Political Economy. *The Annals of the American Academy of Political and Social Science* 568(1) (2000), 41–53.
[24] Patricia Hill Collins. 2004. *Black Sexual Politics: African Americans, Gender, and the New Racism.* Routledge.
[25] Patricia Hill Collins and Sirma Bilge. 2020. *Intersectionality.* Polity Press. https://www.politybooks.com/bookdetail?book_slug=intersectionality-2nd-edition--9781509539673
[26] Brittney Cooper. 2016. Intersectionality. In *The Oxford Handbook of Feminist Theory*, Lisa Disch and Mary Hawkesworth (Eds.). Oxford University Press.
[27] Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum* (1989), 139–168.
[28] Nilanjana Dasgupta and Luis M. Rivera. 2008. When Social Context Matters: The Influence of Long–Term Contact and Short–Term Exposure to Admired Outgroup Members on Implicit Attitudes and Behavioral Intentions. *Social Cognition* 26, 1 (2008), 112–123. https://doi.org/10.1521/soco.2008.26.1.112 arXiv:https://doi.org/10.1521/soco.2008.26.1.112
[29] Jeffrey Dastin. 2019. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. https://reut.rs/2N1dzRJ.
[30] Meera E Deo. 2017. Intersectional Barriers to Tenure. *UCDL Rev.* 51 (2017), 997.
[31] Eva Derous and Roland Pepermans. 2019. Gender Discrimination in Hiring: Intersectional Effects With Ethnicity and Cognitive Job Demands. *Archives of Scientific Psychology* 7, 1 (2019), 40.
[32] Drew DeSilver. 2019. A Majority of U.S. Colleges Admit Most Students Who Apply. *Pew Research Center* (April 2019). https://www.pewresearch.org/fact-tank/2019/04/09/a-majority-of-u-s-colleges-admit-most-students-who-apply/.
[33] Neil J. Dorans. 2002. The Recentering of SAT® Scales and Its Effects on Score Distributions and Score Interpretations. *ETS Research Report Series* 2002, 1 (2002), i–21.
[34] Juliet U. Elu and Linda Loubert. 2013. Earnings Inequality and the Intersectionality of Gender and Ethnicity in Sub-Saharan Africa: The Case of Tanzanian Manufacturing. *American Economic Review* 103, 3 (2013), 289–92.
[35] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. 2022. On Fair Selection in the Presence of Implicit and Differential Variance. *Artificial Intelligence* 302 (2022), 103609. https://doi.org/10.1016/j.artint.2021.103609
[36] Cynthia Fuchs Epstein. 1973. Black and Female-Double Whammy. *Psychology Today* 7(3) (1973), 57–61.
[37] Yuri Faenza, Swati Gupta, and Xuan Zhang. 2020. Impact of Bias on School Admissions and Targeted Intervention. *arXiv preprint arXiv:2004.10846* (2020).
[38] C. FitzGerald, A. Martin, D. Berner, and S. Hurst. 2019. Interventions Designed to Reduce Implicit Prejudices and Implicit Stereotypes in Real World Contexts: A Systematic Review. *BMC Psychol* 7, 1 (May 2019), 29.
[39] A. Gawande. 2010. *The Checklist Manifesto: How to Get Things Right.* Henry Holt and Company. https://books.google.com/books?id=x3IcNujwHxcC
[40] Tamar Szabó Gendler. 2011. On the Epistemic Costs of Implicit Bias. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 156, 1 (2011), 33–63. http://www.jstor.org/stable/41487720
[41] Seth Gershenson, Stephen B. Holt, and Nicholas W. Papageorge. 2016. Who Believes in Me? The Effect of Student–Teacher Demographic Match on Teacher Expectations. *Economics of Education Review* 52 (2016), 209–224. https://doi.org/10.1016/j.econedurev.2016.03.002
[42] Claudia Goldin and Cecilia Rouse. 2000. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review* 90, 4 (September 2000), 715–741. https://doi.org/10.1257/aer.90.4.715
[43] Alexander R. Green, Dana R. Carney, Daniel J. Pallin, Long H. Ngo, Kristal L. Raymond, Lisa I. Iezzoni, and Mahzarin R. Banaji. 2007. Implicit Bias Among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients. *Journal of General Internal Medicine* 22, 9 (2007), 1231–1238.
[44] Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological review* 102, 1 (1995), 4.
[45] Anthony G. Greenwald and Linda Hamilton Krieger. 2006. Implicit bias: Scientific foundations. *California Law Review* 94, 4 (2006), 945–967.
[46] Anthony G. Greenwald and Calvin K. Lai. 2020. Implicit Social Cognition. *Annual Review of Psychology* 71 (2020), 419–445.
[47] Martie G. Haselton and David M. Buss. 2009. Error Management Theory and the Evolution of Misbeliefs. *Behavioral and Brain Sciences* 32, 6 (2009), 522–523. https://doi.org/10.1017/S0140525X09991440
[48] Harry Holzer and David Neumark. 2000. Assessing Affirmative Action. *Journal of Economic Literature* 38(3) (2000), 483–568.
[49] 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
[50] Quinn Capers IV, Daniel Clinchot, Leon McDougle, and Anthony G. Greenwald. 2017. Implicit Racial Bias in Medical School Admissions. *Academic Medicine* 92, 3 (2017), 365–369.
[51] Nathalia Jimenez, Kristy Seidel, Lynn D. Martin, Frederick P. Rivara, and Anne M. Lynn. 2010. Perioperative Analgesic Treatment in Latino and non-Latino Pediatric Patients. *Journal of Health Care for the Poor and Underserved* 21, 1 (2010), 229–236.

[52] Christine Jolls and Cass R Sunstein. 2006. The Law of Implicit Bias. *Calif. L. Rev.* 94 (2006), 969.

[53] Jerry Kang, Mark Bennett, Devon Carbado, Pam Casey, and Justin Levinson. 2011. Implicit Bias in the Courtroom. *UCLa L. rev.* 59 (2011), 1124.

[54] Deborah K. King. 1988. Multiple Jeopardy, Multiple Consciousness: The Context of a Black Feminist Ideology. *Signs* 14, 1 (1988), 42–72. http://www.jstor.org/stable/3174661

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

[56] Arie W Kruglanski and Tallie Freund. 1983. The Freezing and Unfreezing of Lay-Inferences: Effects on Impressional Primacy, Ethnic Stereotyping, and Numerical Anchoring. *Journal of Experimental Social Psychology* 19, 5 (1983), 448–468. https://doi.org/10.1016/0022-1031(83)90022-7

[57] Robert Kurzban and Mark R Leary. 2001. Evolutionary Origins of Stigmatization: The Functions of Social Exclusion. *Psychological bulletin* 127, 2 (2001), 187.

[58] Jennifer S Lerner and Philip E Tetlock. 1999. Accounting for the Effects of Accountability . *Psychological bulletin* 125, 2 (1999), 255.

[59] 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.

[60] 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.

[61] John B. McConahay, Betty B. Hardee, and Batts Valerie. 1981. Has Racism Declined in America? It Depends on Who Is Asking and What Is Asked. *Journal of conflict resolution* 25(4) (1981), 563–579.

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

[63] 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. https://doi.org/10.1073/pnas.1211286109 arXiv:https://www.pnas.org/content/109/41/16474.full.pdf

[64] 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).

[65] Jason A Okonofua and Jennifer L Eberhardt. 2015. Two Strikes: Race and the Disciplining of Young Students. *Psychological science* 26, 5 (2015), 617–624.

[66] 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.

[67] 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.

[68] Julie R. Posselt. 2016. *Inside Graduate Admissions: Merit, Diversity, and Faculty Gatekeeping*. Harvard University Press, Cambridge, MA.

[69] Valerie Purdie-Vaughns and Richard P. Eibach. 2008. Intersectional Invisibility: The Distinctive Advantages and Disadvantages of Multiple Subordinate-Group Identities. *Sex Roles* 59(5-6) (2008), 377–391.

[70] Barbara F. Reskin and Debra Branch McBrier. 2000. Why Not Ascription? Organizations' Employment of Male and Female Managers. *American Sociological Review* 65, 2 (2000), 210–233. http://www.jstor.org/stable/2657438

[71] Dan-Olof Rooth. 2010. Automatic Associations and Discrimination in Hiring: Real World Evidence. *Labour Economics* 17, 3 (2010), 523–534.

[72] Ashleigh Shelby Rosette and Robert W. Livingston. 2012. Failure Is Not an Option for Black Women: Effects of Organizational Performance on Leaders With Single Versus Dual-Subordinate Identities . *Journal of Experimental Social Psychology* 48, 5 (2012), 1162–1167. https://doi.org/10.1016/j.jesp.2012.05.002

[73] 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.

[74] Howard Schuman, Charlotte Steeh, Lawrence Bobo, and Maria Krysan. 1985. *Racial Attitudes in America: Trends and Interpretations*. Harvard University Press. https://books.google.com/books?id=EhfhDwva0VYC

[75] 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/

[76] Amartya Sen. 2006. *Identity and Violence: The Illusion of Destiny*. W.W. Norton, New York, NY.

[77] Hege Skjeie. 2015. Gender Equality and Nondiscrimination: How to Tackle Multiple Discrimination Effectively. In *Visions for Gender Equality*, Francesca Bettio and Silvia Sansonetti (Eds.). European Union, Luxembourg, 79–82.

[78] Thomas Sowell. 2004. *Affirmative Action Around the World: An Empirical Study*. Yale University Press.

[79] Cheryl Staats. 2016. Understanding Implicit Bias: What Educators Should Know. *American Educator* 39, 4 (2016), 29.

[80] Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *NeurIPS*. 13209–13220.

[81] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131.

[82] 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.

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

[84] Christine Wennerås and Agnes Wold. 1997. Nepotism and Sexism in Peer-Review. *Nature* 387, 6631 (01 May 1997), 341–343. https://doi.org/10.1038/387341a0

[85] 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.

[86] 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.

[87] Jonathan C. Ziegert and Paul J. Hanges. 2005. Employment Discrimination: The Role of Implicit Attitudes, Motivation, and a Climate for Racial Bias. *Journal of Applied Psychology* 90, 3 (2005), 553–562.