The Effect of the Rooney Rule on Implicit Bias in the Long Term

L. Elisa Celis
Yale University

Chris Hays
Yale University

Anay Mehrotra
Yale University

Nisheeth K. Vishnoi
Yale University

ABSTRACT

The Rooney Rule, originally proposed to counter implicit bias in hiring, has been implemented in the private and public sector in various settings. This rule requires that a decision-maker include at least one candidate from an underrepresented group in their shortlist of candidates. Recently, [42] proposed a mathematical model of implicit bias and studied the effectiveness of the Rooney Rule when applied to a single selection decision. However, selection decisions often occur repeatedly over time; e.g., a software firm is continuously hiring employees or a university makes admissions decisions every year. Further, it has been observed that, given consistent counterstereotypical feedback, implicit biases against underrepresented candidates can change.

In this paper, we propose a model of how a decision-maker's implicit bias changes over time given their hiring decisions either with or without the Rooney Rule in place. Our model draws from the work of [42] and the literature on opinion dynamics. Our main result is that, for this model, when the decision-maker is constrained by the Rooney Rule, their implicit bias roughly reduces at a rate that is inverse of the size of the shortlist—independent of the total number of candidates, whereas without the Rooney Rule, the rate is inversely proportional to the number of candidates. Thus, our model predicts that when the number of candidates is much larger than the size of the shortlist, the Rooney Rule enables a significantly faster reduction in implicit bias, providing additional reason in favor of instating it as a strategy to mitigate implicit bias. Towards empirically evaluating the long-term effect of the Rooney Rule in repeated selection decisions, we conduct an iterative candidate selection experiment on Amazon Mechanical Turk. We observe that, indeed, decision-makers subject to the Rooney Rule select more minority candidates *in addition to* those required by the rule itself than they would if no rule is in effect, and in fact are able to do so without considerably decreasing the utility of candidates selected.

ACM Reference Format:
L. Elisa Celis, Chris Hays, Anay Mehrotra, and Nisheeth K. Vishnoi. 2021. The Effect of the Rooney Rule on Implicit Bias in the Long Term. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 31 pages. https://doi.org/10.1145/3442188.3445930

FAccT '21, March 3–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03…$15.00
https://doi.org/10.1145/3442188.3445930

1 INTRODUCTION[⋆]

Implicit bias is the unconscious association of certain qualities (or lack thereof) to individuals of socially salient groups, like those defined by race, gender, or sexuality. In recent decades, a large body of experimental research has demonstrated the adverse effects of implicit bias in a wide range of contexts, including hiring [49, 58, 68], university admissions [35, 55], criminal justice [6, 32, 40], and healthcare [15, 29, 37]. In fact, even when decisions are based on quantifiable characteristics of the applicants, decision-makers can systematically undervalue underrepresented candidates. For instance, it was found that women in managerial positions had to show roughly twice as much evidence to be seen as equally competent to men [48, 66], and evaluators, across jobs, unknowingly customized their evaluation criteria to favor the stereotypical gender [63].

The Rooney Rule is a simple and widely adopted policy to counteract the adverse effects of implicit bias [19, 57, 64]. It requires that at least one among a shortlist of candidates (for further interviews or evaluation) picked by a decision-maker must come from an underrepresented group. It was originally instituted in the interview processes for hiring head coaches in the NFL in 2003, and since then, has been adopted by various corporations such as Amazon, Facebook, and Microsoft [52], and in several public sector contexts in the US [9, 26]. In fact, in 2020, the NFL strengthened the Rooney Rule to require at that least two minority candidates in the shortlist of interviewees for head coaching positions be from an underrepresented group [67]. This motivates a generalization of the rule to the ℓ-th order Rooney Rule (also proposed in [42]), which requires at least ℓ of the shortlisted candidates to be in the underrepresented group.

Although there is evidence that the Rooney Rule has had a positive impact in various contexts [19, 25], it is a subject of much debate [19, 64]. Proponents of the policy argue that it counteracts the effects of implicit bias, while critics warn that it can lead to poorer selections.

Towards demonstrating the effectiveness of the Rooney Rule, [42] proposed a mathematical model of implicit bias and showed that the Rooney rule can improve the "true" utility of the decision-maker in a single hiring decision. More precisely, they consider n candidates partitioned into two disjoint groups $G_X, G_Y \subseteq [n]$, where G_X is the group of underrepresented candidates. Each candidate has a true, *latent utility*, which is the value they would contribute if selected, and an *observed utility* which is the decision-maker's (potentially biased) estimate of their latent utility. They model the decision-maker's implicit bias as a multiplicative factor $\beta \in [0, 1]$, such that the observed utility of underrepresented

[⋆] The code for the experiment in Section 5 is available at [11]. To complete a version of the experiment similar to the one presented to participants, see [12].

candidates (those in G_X) is β times their latent utility, while the observed utility of all other candidates (those in G_Y) is the same as their latent utility.[1] Thus, if $\beta = 1$, the decision-maker evaluates underrepresented candidates without bias, and its bias against them becomes more severe as β approaches 0. The decision-maker shortlists k candidates (out of n) with the highest observed utility. In the setting where n is much larger than k, [42] characterize conditions on β, the proportion of underrepresented candidates ($\rho := \frac{|G_X|}{n}$), and the distribution of latent utilities, such that under these conditions, applying the Rooney Rule (for $\ell = 1$) increases the total latent utility of the shortlisted candidates.

An important benefit of the Rooney Rule is that the decision-maker has the opportunity to closely evaluate qualified underrepresented candidates, see that their latent utility was greater than expected and learn to evaluate underrepresented candidates more accurately. Indeed, studies show that implicit biases can change over time [16, 61] and with changes in local-environments [21]. In particular, it has been observed that exposure to other groups [3, 23] and counterstereotypical evidence opposing the implicit beliefs [8, 22] can help reduce implicit bias. Thus, one would hope that as the decision-maker observes the latent utilities of more underrepresented candidates over multiple iterations of selection, its implicit biases would change. This is in line with work on belief and opinion formation, which model how individuals update their beliefs and opinions based on the information they observe [2, 39]. At a high-level, these works model the beliefs of individuals using probability distributions, where, each time an individual receives new information, the distribution is updated to incorporate the new information and reflect the corresponding new beliefs [18, 36].

1.1 Our contributions

We consider a mathematical model for implicit bias and how it updates each selection decision. Under the assumptions of the model, the Rooney Rule provably enables a significantly faster reduction in implicit bias of the decision-maker over multiple selection decisions when compared to the unconstrained condition; this gives a mathematical explanation for the aforementioned empirical observations.

Our model maintains a probability distribution over the implicit bias of the decision-maker and updates this distribution after each iteration depending on the ratio of the latent utility of the shortlisted candidates and the observed utility of the shortlisted candidates; see Section 2. Technically, we show that, when the decision-maker uses the ℓ-th order Rooney Rule for $\ell \geq 1$, its implicit bias reduces, roughly, at the rate of $\frac{1}{(k-\ell+1)}$ – independent of n (Theorem 3.1), whereas, when the decision-maker is not constrained by the Rooney Rule, then the rate at which its implicit bias reduces is, roughly, $\frac{1}{n}$ (Theorem 3.2). Thus, when the number of applicants n is much larger than the size of the shortlist k, our model predicts that using the Rooney Rule leads to a significantly faster reduction in the decision-maker's implicit bias. Expanding on these results, we characterize the effect of other parameters (such as the proportion of underrepresented candidates ρ) on the change in the decision-maker's implicit bias over time (Section 3.1). We also discuss how our results generalize to other models, where the decision-maker's implicit bias is drawn from distributions not in the beta family and updated using other rules (Section 3.2). Thus, our theoretical results complement the work of [42] and provide an additional reason to instate the Rooney Rule as a strategy to mitigate implicit bias.

Towards empirically evaluating the effect of the Rooney Rule in repeated selections, we enlist participants on Amazon's Mechanical Turk to participate in an iterative selection experiment (Section 5). We represented candidates from two different groups with different colored tiles and applied bias to the observed utilities of one of the groups. In each iteration, participants were incentivized to maximize the latent utility of their selection, and the latent utilities of their selections were revealed after each round. We observe that the participants subject to the Rooney Rule selected significantly more underrepresented candidates *in addition to* those required by the rule itself than participants not subject to the rule, without substantially decreasing the utility of candidates selected.

1.2 Related work

Implicit bias. Studying implicit bias is a rich field in psychology [30, 31] and several works study the origins of implicit bias [53, 59], its adverse effects [48, 60, 66], and its long-term trends [16]. We point the reader to the excellent treatise [41] for an overview of the field.

[42] introduce a model for implicit bias, and under this model, characterized conditions where the Rooney Rule improves the latent utility of the selection. Under the same model, [14] study the ranking problem (a generalization of selection) under implicit bias, and propose simple constraints on rankings which improve the latent utility of the output ranking. Both [42] and [14] consider the latent utility in a single instance of the problem, whereas, we are interested in how the implicit bias of the decision-maker changes over multiple iterations.

[28] study selection under a different model of bias: where the decision-maker's observed utility has higher than average noise for underrepresented candidates. They consider a family of constraints, and show that, in their model, these constraints always increase the latent utility. Unlike them, our goal is to understand the effects of constraints on the implicit bias of the decision-maker. In Section 3.2, we discuss how our results generalize when there is noise in $\beta^{(t)}$. Accounting for other forms of noise in the observed utilities can be an interesting extension to this work.

Belief update models. Works on opinion dynamics and social learning study mathematical models of how people's beliefs change when they gain new information [1, 18, 36]. Several works in this field represent beliefs by probability distributions and study simple rules, similar to the one we consider, to update these distributions [18, 24]. We refer the reader to [2] for a comprehensive overview of the field.

In a similar vein, the theory of subjective logic [38] mathematically models beliefs under uncertainty. A seminal work [39] gives a mapping from beliefs to a beta distribution Beta(a, b); roughly, a is the evidence favouring the belief and b evidence against it. A canonical example is the statement: "A ball drawn at random (from an urn of red and black balls) will be red" [39]. A person observes

[1] To be precise, [42] consider $\beta \in (1, \infty)$, and assume that the observed utility of an underrepresented candidate is $1/\beta$ times their latent utility. Considering $\beta \in [0, 1]$ is more convenient in our setting.

multiple draws from the urn, and after each draw updates their belief. Our model uses the beta distribution to model the decision-maker's implicit bias parameter and, in relevant contexts, can be viewed as the decision-maker's "belief" in the following statement: *the latent utilities of candidates from group G_X and group G_Y are identically distributed.*

Long-term impact. Several prior works have studied the long-term impacts of affirmative action policies on society [34, 46, 51]. [46] consider how common fairness constraints in classification settings affect the underlying population over time. [51] assumes individuals have binary utilities, (qualified or unqualified), and they give asymptotic results for a broad-set of dynamics depending on how the fraction of qualified individuals in each group changes. [34] studies a dynamics in the context of a labor market.

In contrast, we allow for non-binary utilities, study the effect on the decision-maker's implicit bias, and give non-asymptotic results.

Iterated learning experiments. In a classic formulation of a *function learning experiment*, in each iteration, participants are given a numeric input and asked to predict its numeric output from some examples or using knowledge they accumulated so far. Several experiments in cognitive science [10, 43, 47] and behavioral economics [20, 33] use function learning experiments to study human performance on prediction tasks with incomplete information. Related to our work, these techniques have also been used to (attempt to) measure implicit biases [45]. We refer the reader to [10] for an overview of the experimental work in this topic.

Our empirical experiment builds on the classic iterated learning experimental design—there is a simple linear relationship between the observed (input) and latent utilities (output). However, we do not ask for participants to explicitly predict the output; instead, they implicitly do so by selecting the observed utilities which they they predict will have the highest latent utilities (output).

Notation

For a natural number $n \in \mathbb{N}$ by $[n]$ we denote the set $\{1, 2, \ldots, n\}$, and for a real number $x \in \mathbb{R}$ by $\exp(x)$ we denote e^x. We use calligraphic letters such as $\mathcal{D}$ and $\mathcal{P}$ to denote *distributions*, and $X \sim \mathcal{D}$ denotes a sample X drawn from $\mathcal{D}$. For a distribution $\mathcal{D}$ with probability density function f, its support is the set

$$\{x \colon f(x) > 0\},$$

we denote the support of $\mathcal{D}$ by $\mathrm{supp}(\mathcal{D})$. If the support of a distribution is an interval over the reals, we say that the distribution is continuous. We use (t) in the superscript to indicate the t-th iteration. We use the subscript i to index the underrepresented candidates and j to index all other candidates. We use $\mathrm{Beta}(a, b)$ to denote the beta distribution with parameters a and b. It holds that

$$\mathbb{E}_{\beta \sim \mathrm{Beta}(a,b)}[\beta] = \frac{a}{(a+b)}.$$

Formally, given $a, b \geq 0$, define $\mathrm{Beta}(a, b)$ to be the distribution with the following cumulative density function: for all $x \in [0, 1]$

$$\Pr_{\beta \sim \mathrm{Beta}(a,b)}[\beta \leq x] := \frac{\int_0^x y^{a-1}(1-y)^{b-1}dy}{\int_0^1 y^{a-1}(1-y)^{b-1}dy}.$$

2 MODEL

In each round of selection, there are n candidates, and a decision-maker shortlists k of them. The candidates are partitioned into two disjoint groups $G_X, G_Y \subseteq [n]$, where G_X denotes the group of underrepresented candidates. The intersection $G_X \cap G_Y$ is empty and $G_X \cup G_Y = [n]$. We call the candidates in G_X the *X-candidates* and those in G_Y the *Y-candidates*. Each candidate has a true or *latent utility* which is the value which a candidate would contribute if selected. Denote this utility by $X_i \geq 0$ for the i-th X-candidate, and by $Y_j \geq 0$ for the j-th Y-candidate. We assume that the latent utilities of all candidates are independently and identically (i.i.d.) drawn in each iteration from some continuous distribution $\mathcal{P}$. We assume that $\mathcal{P}$ has non-negative and bounded support. Let ρ be the fraction of the underrepresented candidates:

$$\rho := \frac{|G_X|}{|G_X| + |G_Y|}.$$

While the utilities of individual candidates change with time, we assume that ρ itself does not change and, as a consequence, under our assumption G_X and G_Y also do not change.

2.1 Implicit bias model of [42]

In [42], based on the empirical observations of [65], the setting where the decision-maker does not observe latent utilities and instead sees an *observed utility*, which is its (possibly biased) estimate of the latent utilities is considered. They consider the following model of observed utilities parameterized by an implicit bias parameter $\beta \in [0, 1]$: define the observed utilities of an X-candidate $i \in G_X$ as

$$\widetilde{X}_i := \beta \cdot X_i.$$

The observed utility of a Y-candidate is assumed to be the same as its latent utility. Notice that in the above definition, if $\beta = 1$, then $\widetilde{X}_i = X_i$ for all $i \in G_X$ and the decision-maker evaluates X-candidates without bias. It is sometimes useful to define the following vectors: $X := (\ldots, X_i, \ldots)$, $\widetilde{X} := (\ldots, \widetilde{X}_i, \ldots)$ and $Y := (\ldots, Y_j, \ldots)$, where i varies over G_X and j varies over G_Y.

2.2 Candidate selection problems and the Rooney Rule

The utility of a subset of candidates S is defined as the sum of the utilities of all candidates in S. Given a subset $S \subseteq [n]$, define its *total observed utility* as

$$\textsc{Util}(S, \widetilde{X}, Y) := \sum_{i \in S \cap G_X} \widetilde{X}_i + \sum_{j \in S \cap G_Y} Y_i. \tag{1}$$

Similarly, define the *total latent utility* of $S \subseteq [n]$ as the sum of the latent utilities of all candidates in S: $\textsc{Util}(S, X, Y)$ (where we replace $\widetilde{X}$ in Equation (1) by X).

As in [42], we assume that the decision-maker selects a subset of candidates S of size k which maximizes total observed utility:

$$S := \operatorname*{argmax}_{T \subseteq [n] \colon |T| = k} \textsc{Util}(T, \widetilde{X}, Y). \tag{2}$$

Note that when $\beta < 1$, the set S may have few X-candidates, disadvantaging those candidates.

The ℓ-th order Rooney Rule tries to address this by requiring the decision-maker to select at least ℓ X-candidates. (Note that if $\ell = 1$, then the ℓ-th order Rooney Rule is the same as the usual Rooney Rule). Let

$$\mathcal{R}(\ell) := \{T \subseteq [n] \colon |T \cap G_X| \geq \ell \text{ and } |T| = k\} \tag{3}$$

be the set of all subsets of size k satisfying the ℓ-th order Rooney Rule. The decision-maker constrained by the ℓ-th order Rooney Rule picks a subset $S_\ell \in \mathcal{R}(\ell)$ satisfying the rule which maximizes the total observed utility:

$$S_\ell := \operatorname*{argmax}_{T \in \mathcal{R}(\ell)} \textsc{Util}(T, \widetilde{X}, Y). \tag{4}$$

Notice that the set S selected by the decision-maker without the the Rooney Rule (in Equation (2)) is the same as the set S_0 above.

As we discuss next, the decision-maker updates its beliefs as a function of the total latent utility and the total observed utility of S_ℓ. To simplify the notation there, let the U and $\widetilde{U}$ be the total latent utility and the total observed utility of S_ℓ:

$$U := \textsc{Util}(S_\ell, X, Y), \tag{5}$$

$$\widetilde{U} := \textsc{Util}(S_\ell, \widetilde{X}, Y). \tag{6}$$

Note that, before selecting the candidates, the decision-maker knows $\widetilde{U}$, but does not know U. Our implicit bias update model assumes that the decision-maker sees U *after* selecting the candidates. The rationale is that after the decision-maker selects the candidates $S_\ell \subseteq [n]$ and observes their actual performance—at the job or in an interview—it can better estimate their latent utility U.

2.3 Implicit bias update model

Inspired by the works [39, 62] that give mappings from beliefs to beta distributions, in each iteration, we model the implicit bias β as a draw from the a beta distribution $\text{Beta}(a, b)$, where $a, b > 1$ and, roughly, a is the evidence favouring the belief and b evidence against it.

In our setting, a is roughly proportional to the belief that X-candidates have the same latent utility as Y-candidates, and b is roughly proportional to bias against it. Notice that the larger a is, the closer $\mathbb{E}[\beta]$ is to 1 (no bias), and the larger b is, the closer $\mathbb{E}[\beta]$ is to 0 (largest bias).[2]

If $U > \widetilde{U}$, then the decision-maker has evidence that the X-candidates performed better than expected. In this case, the decision-maker's implicit bias reduces, i.e., a would increase. Since $\beta \in [0, 1]$, we can argue that $U \geq \widetilde{U}$. To see this, note that

$$(U - \widetilde{U}) = (1 - \beta) \cdot \sum_{i \in S_\ell \cap G_X} X_i,$$

and since $1 - \beta \geq 0$ and $X_i \geq 0$ for all i, we have that

$$(U - \widetilde{U}) \geq 0.$$

Thus, the decision-maker does not receive evidence to support it's bias, and so, b is a constant in this model.

To summarize, given parameter $a > 1$ which varies over iterations and some fixed parameter $b > 1$, we consider the distribution

$$\mathcal{D}(a) := \text{Beta}(a, b),$$

and draw the decision-maker's implicit bias β from $\mathcal{D}(a)$.

Since we consider multiple iterations of the above model of candidate selection, we need to specify how β evolves. Let a superscript (t) on a variable indicate the variable's value at the t-th iteration. We start with $a^{(1)}$ to be some fixed number greater than 1. Suppose that in the t-th iteration, the decision-maker selects a subset $S_\ell^{(t)}$ whose total latent utility is $U^{(t)}$ and total observed utility is $\widetilde{U}^{(t)}$. We propose and study the following update rule:

$$a^{(t+1)} := \frac{U^{(t)}}{\widetilde{U}^{(t)}} \cdot a^{(t)}, \tag{Update rule; 7}$$

$$\beta^{(t+1)} \sim \mathcal{D}(a^{(t+1)}).$$

We summarize the complete mathematical model in Model 1.

Model 1: Our implicit bias update model

```
Initialize: The Rooney Rule parameter $\ell \in \mathbb{Z}_{\geq 0}$, a parameter
            $a^{(1)} > 1$, and a constant $b > 1$.
For $t = 1, 2, \ldots$ do
    Sample $\{Y_j^{(t)}\}_{j \in G_Y}$ and $\{X_i^{(t)}\}_{i \in G_X}$ i.i.d. from $\mathcal{P}$.
    Sample $\beta^{(t)}$ from $\mathcal{D}(a^{(t)}) := \text{Beta}(a^{(t)}, b)$.          //Implicit bias
    Define $\widetilde{X}_i^{(t)} := \beta^{(t)} \cdot X_i^{(t)}$ for all $i \in G_X$.          //Observed utilities
    Select $S_\ell^{(t)} := \text{argmax}_{S \in \mathcal{R}(\ell)} \textsc{Util}(S, \widetilde{X}^{(t)}, Y^{(t)})$.          //Shortlist
    Let $U^{(t)} := \textsc{Util}(S_\ell^{(t)}, X, Y)$ and $\widetilde{U}^{(t)} := \textsc{Util}(S_\ell^{(t)}, \widetilde{X}, Y)$.
    Update $a^{(t+1)} := \left(\frac{U^{(t)}}{\widetilde{U}^{(t)}}\right) \cdot a^{(t)}$.          //Update implicit bias
```

2.4 Discussion of the model

While we consider a simple multiplicative update rule, one could also study other update rules. For instance, some prior works on opinion dynamics study additive updates and updates where the new value is a convex combination of the current value and observation [24, 44]. We discuss how our results generalize to other update rules in Section 3.2.2.

Note that our model assumes that the decision-maker evaluates the latent utility of all selected candidates together, and so observes $U^{(t)}$ and $\widetilde{U}^{(t)}$ and not the latent utilities of individual candidates; we discuss other scenarios in Section 6.

The update rule (7) in our model is only a function of the current belief (i.e., $\mathcal{D}^{(t)}$) and observation (i.e., $\frac{U^{(t)}}{\widetilde{U}^{(t)}}$). This is equivalent to the assumption made by the Bayesian-Without-Recall model [18, 56]. (Note, however, that the update in Equation (7) is not Bayesian). Our update rule also bears resemblance to the multiplicative weights update method [4] that has been successfully used to explain learning in social groups [13] and sexual evolution [17].

While our update rule does encode past observations, considering more complicated rules which explicitly include all past observations can lead to interesting extensions of this work.

Our model incorporates uncertainty in the decision-maker's implicit bias at each iteration by drawing from a distribution, but assumes that the utilities do not have any noise. It would be interesting to consider models which incorporate noise in the observations of the committee as well.

[2] This follows since the expected value of a $\text{Beta}(a,\ b)$ random variable is $a/(a+b)$.

3 THEORETICAL RESULTS

Our first result considers the case when the decision-maker uses the Rooney Rule. It shows that, under the implicit bias update rule (1), when the decision-maker uses the Rooney Rule, the reduction in its implicit bias is independent of n.

Theorem 3.1 (**Fast learning with the Rooney Rule**). *Under the implicit bias update rule* (1)*, given parameters* $a^{(1)}, b > 1$ *there exists a constant* $C_1 > 0$*, such that, for all iterations* $t \in \mathbb{N}$*, population size* $n \in \mathbb{N}$*, number of candidates selected* $k \in [n]$*, parameter* $\ell \in [k]$*, ratio of the underrepresented candidates* $\rho \in (0,1)$*, and continuous and bounded distribution of latent utility* $\mathcal{P}$*, when the decision-maker is constrained by the* ℓ*-th order Rooney Rule, then*

$$\mathbb{E}\left[\beta^{(t)}\right] \geq \left(\frac{C_1 \cdot \frac{t\rho}{(k-\ell+1)}}{1 + C_1 \cdot \frac{t\rho}{(k-\ell+1)}}\right) \cdot \left(1 - e^{-t\rho/16}\right) \tag{8}$$

where the expectation is over the draws of implicit bias $\beta^{(s)}$ *and latent utilities of candidates in all previous iterations* $s \in [t-1]$.

Note that the lower bound in Equation (8) is independent of n and, very roughly, scales as $\frac{\rho}{k-\ell+1}$. At a high-level, this is because the Rooney Rule ensures that the decision-maker selects at least $\frac{\ell}{k}$ fraction of the shortlist from G_X, and this fraction does not vary with n. As we show in our next result, this independence doesn't hold if the decision-maker does not use the Rooney Rule ($\ell = 0$).

Theorem 3.1 also shows that if the decision-maker uses the ℓ-th order Rooney Rule, in the long term (i.e., for a large t), its expected implicit bias is almost 1 (no bias) independent of the number of candidates $n \in \mathbb{N}$. To see this, let t be a large value compared to $\frac{k-\ell+1}{\rho}$, then notice that both terms containing t in Equation (8) are close to 1 and do not depend on n.

Our next result considers the case when the decision-maker does not use the Rooney Rule. It assumes that $\mathcal{P}$ is the uniform distribution on $[0,1]$ (denoted by Unif$(0,1)$); we discuss how the result extends to other distributions in Section 3.2.3.

Theorem 3.2 (**Slow learning without the Rooney Rule**). *Under the implicit bias update rule* (1)*, given parameters* $a^{(1)}, b > 1$*, for all iterations* $t \in \mathbb{N}$*, number of candidates selected* $k \in [n]$*, and ratio of the underrepresented candidates* $\rho \in (0,1)$*, when the decision-maker is not constrained by the Rooney Rule (i.e.,* $\ell = 0$*) and* $\mathcal{P}$ *is* Unif$(0,1)$*, there exists a* $n_0 \in \mathbb{N}$ *and* $C_2 > 0$*, such that for all* $n \geq n_0$

$$\mathbb{E}\left[\beta^{(t)}\right] \leq \mathbb{E}\left[\beta^{(1)}\right] + C_2 \cdot \frac{t \ln n}{n(1-\rho)}, \tag{9}$$

where the expectation is over draws of implicit bias $\beta^{(s)}$ *and latent utilities of candidates in all previous iterations* $s \in [t-1]$.

Thus, when the decision-maker is not constrained by the Rooney Rule, the rate at which its implicit bias reduces (approaches 1) is no more than, roughly, $\frac{\ln n}{n(1-\rho)}$. To see this, observe that rewriting Equation (9), we get $\frac{1}{t} \cdot \mathbb{E}[\beta^{(t)} - \beta^{(1)}] \leq C_2 \cdot \frac{\ln n}{n(1-\rho)}$. Further, from Theorem 3.2 we can infer that for any fixed $t \in \mathbb{N}$ there is a large enough n, such that decision-maker's expected implicit bias after t iterations is almost the same as its initial implicit bias: $\mathbb{E}[\beta^{(1)}]$.

Thus, together with Theorem 3.1, Theorem 3.2 shows a significant difference in the decision-maker's implicit bias at a given iteration with and without the Rooney Rule when n is much larger than k, providing evidence to support implementing the Rooney Rule. We note that Theorems 3.1 and 3.2 hold for any value of $t = 1, 2, \ldots$; and not just in the limit when t is large. The upper bound in Theorem 3.2 decreases (becomes weaker) as t increases. This is necessary as we can show that for a fixed n as $t \to \infty$, the decision-maker becomes unbiased (see Lemma B.8 for a formal statement).

Finally, we remark that in the statements of our theorems, we have tried to capture the dependence on each parameter as cleanly as possible and not tried to optimize the constants.

3.1 Qualitative predictions

We now summarize the qualitative effects of the parameters of our model on the decision-maker's expected implicit bias.

3.1.1 Effect of n. As discussed above, when the decision-maker does not use the Rooney Rule, the rate at which its implicit bias decreases slows down with n (Theorem 3.2):

$$\ell = 0: \qquad \text{bias reduction rate} \propto \frac{1}{n}.$$

Intuitively, this is because, as the number of available Y-candidates increases, a biased decision-maker selects fewer X-candidates. Thus, the decision-maker becomes less likely to observe that X-candidates perform better than expected.

In contrast, when the decision-maker uses the Rooney Rule, its expected implicit bias is independent of n (Theorem 3.1):

$$\ell \geq 1: \qquad \text{bias reduction rate is independent of } n.$$

3.1.2 Effect of ℓ. When the decision-maker uses the Rooney Rule, increasing ℓ increases its learning (Theorem 3.1):

$$\ell \geq 1: \qquad \text{bias reduction rate} \propto \ell.$$

This is intuitive, as increasing ℓ leads the decision-maker to select a larger number of X-candidates, and hence, is more likely to observe that they perform better than they expected. Mathematically, this is because the update, $\frac{\widetilde{U}^{(t)}}{U^{(t)}}$, is an increasing function of the total utility of the X-candidates selected. And the total utility of the X-candidates increases on selecting more candidates. One might think that this means setting $\ell = k$ is the optimal choice. However, in the real world, reducing the decision-maker's bias is not the only objective—and the choice of ℓ should also consider other factors.

3.1.3 Effect of k. Increasing k, holding everything else fixed, decreases the decision-maker's expected implicit bias at any given iteration (Theorem 3.1):

$$\ell \geq 1: \qquad \text{bias reduction rate} \propto \frac{1}{k}.$$

3.1.4 Effect of ρ. Both the lower bound (8) and upper bound (9) are increasing functions of ρ. We can infer that in both cases increasing ρ (fixing other parameters) increases $\mathbb{E}[\beta^{(t)}]$:

$$\ell \geq 0: \qquad \text{bias reduction rate} \propto \rho.$$

Intuitively, increasing ρ increases the fraction of X-candidates, and makes it more likely for the decision-maker to select an X-candidate as the number of positive outliers (whose observed utility appears to be good *despite* the bias) increases.

3.2 Generalizations of our theoretical results

3.2.1 Generalizing to other distributions of implicit bias. In this section, we show how our results generalize when $\mathcal{D}(a)$ is not a beta distribution. This can be interesting, for instance, if the decision-maker's implicit biases do not follow the beta family, or when the draws of $\beta^{(t)}$ are noisy.

Notation. Let $\mathcal{D}(a)$ be any continuous distribution supported on $[0,1]$ and parameterized by $a > 1$. Define $\Phi(x)$ as the expected value of β drawn from $\mathcal{D}(x)$:

$$\Phi(x) \coloneqq \mathbb{E}_{\beta\sim\mathcal{D}(x)}[\beta],$$

and median(a) as the median of $\beta \sim \mathcal{D}(a)$. For example, we can consider $\mathcal{D}(a)$ to be the truncated normal distribution with mean a and fixed variance, in which case, median$(a) = a$.

Assume that

1. $\Phi(\cdot)$ is an increasing function,
2. $\Phi(\cdot)$ is concave,
3. $\mathbb{E}_{\beta\sim\mathcal{D}(a)}\left[\frac{1}{\beta}\right]$ is decreasing in a and finite for every $a > 1$, and
4. there is a constant $C_3 > 0$, such that, for all $a > 1,\ k \in [n], \ell \in [k]$
$$\frac{1-\mathrm{median}(a)}{\mathrm{median}(a)+(k-\ell)} > \frac{C_3}{a\cdot(k-\ell+1)}.$$

We prove the following versions of Theorem 3.1 and Theorem 3.2:

- *(Informal).* For all $a^{(1)} > 1$ there is a constant $C_4 > 0$, such that for all $t \in \mathbb{N},\ n \in \mathbb{N},\ k \in [n],\ \ell \in [k]$, and continuous and bounded $\mathcal{P}$, when the decision-maker applies the Rooney Rule, the following holds
$$\mathbb{E}[\beta^{(t)}] \geq \Phi\left(a^{(1)} + \frac{C_3}{C_4}\frac{t\rho}{(k-\ell+1)}\right)\cdot\left(1-e^{-\frac{t\rho}{16}}\right). \tag{10}$$
- *(Informal).* For all $a^{(1)} > 1$ there is a constant $C_5 > 0$, such that for all $t \in \mathbb{N},\ k \in \mathbb{N},\ \ell \in [k]$, when the decision-maker applies the Rooney Rule and $\mathcal{P} \coloneqq \mathrm{Unif}(0,1)$, there is a constant $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$ the following holds
$$\mathbb{E}[\beta^{(t)}] \leq \Phi\left(a^{(1)} + C_5\frac{t\ln n}{n(1-\rho)}\right) + \frac{t\ln n}{n(1-\rho)}. \tag{11}$$

(We present their formal statements and proofs in Supplementary Material B.1.)

Discussion of assumptions. In Model (1), a, roughly, represents how unbiased the decision-maker is. Thus, it is natural to expect Assumption (1) and first part of Assumption (3). We can avoid Assumption (2) (we explain this with the proof); although then the upper bound (11) becomes slightly weaker. The second part of Assumption (3), intuitively says that the decision-maker's implicit bias does not have a large probability mass near 0 (extreme bias). Finally, Assumption (4) upper bounds the median(a) in terms of a: it says that the median is not too close to 1 (no bias) for a given a.

Discussion. Allowing for different distributions of implicit bias, potentially due to noise, is related to [28], who consider noise in the decision-maker's implicit bias in one iteration. However, they draw a different value of implicit bias for each candidate and consider different levels of noise for both groups of candidates. Expanding Model 1 to incorporate these would be an interesting direction for future work.

Remark 3.3. *Note that if we substitute the definition of* $\Phi(\cdot)$ *for the beta distribution in Equation* (10) *we recover Theorem 3.1, and if substitute the definition in Equation* (11) *we recover Theorem 3.2.*

3.2.2 Generalizing to other update rules. In this section, we consider the update rule

$$a^{(t+1)} \coloneqq a^{(t)}\cdot F\left(\frac{U^{(t)}}{\widetilde{U}^{(t)}}\right), \tag{12}$$

where $F\colon [1,\infty) \to [1,\infty)$ is a continuous function satisfying some mild assumptions. In particular, we assume that

1. $F(1) = 1$,
2. F is strictly increasing, and
3. F is concave.

We prove the following versions of Theorem 3.1 and Theorem 3.2:

- *(Informal).* Under update rule (12), if F satisfies the above assumptions and the decision-maker uses the Rooney Rule, then for all $\varepsilon \in (0,1)$, there exists an iteration $t \in \mathbb{N}$, such that,
$$\mathbb{E}[\beta^{(t)}] \geq 1-\varepsilon.$$
- *(Informal).* Under update rule (12), if F satisfies the above assumptions and the decision-maker does not use the Rooney Rule, then for all $\varepsilon \in (0,1)$ and $t \in \mathbb{N}$, there exists an $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$
$$\mathbb{E}[\beta^{(t)}] \leq \mathbb{E}[\beta^{(1)}] + \varepsilon.$$

(We present their formal statements and proofs Supplementary Material B.2.)

3.2.3 Generalizing Theorem 3.2 to other distributions. We can extend Theorem 3.2 to other continuous and bounded distributions of latent utility $\mathcal{P}$. In this case, we need to make a mild assumption on $\mathcal{P}$, which is satisfied by several distributions.

Notation. Let M be the supremum of the support of $\mathcal{P}$: $M \coloneqq \sup(\mathrm{supp}(\mathcal{P}))$. Define $T_{\mathcal{P}}(\varepsilon)$ as the probability that an X drawn from $\mathcal{P}$ is at least $(1-\varepsilon)M$. Formally, $T_{\mathcal{P}}(\varepsilon) \coloneqq \Pr_{X\sim\mathcal{P}}[X \geq (1-\varepsilon)M]$.

Assume that: there exists a function $\varepsilon\colon \mathbb{N} \to \mathbb{R}_{>0}$ satisfying

$$\varepsilon(n) + e^{-n\cdot T_{\mathcal{P}}(\varepsilon(n))} \leq \frac{1}{\mathrm{poly}(n)}. \tag{Assumption; 13}$$

Then we get the following upper bound in Theorem 3.2:

$$\mathbb{E}\left[\beta^{(t)}\right] \leq \mathbb{E}\left[\beta^{(1)}\right] + \frac{t}{\mathrm{poly}(n)}.$$

Here, poly$(\cdot)$ hides the dependence on $(1-\rho)$. We do not state the dependence on $(1-\rho)$ as it changes with the choice of $\mathcal{P}$.
(We present the formal statement as Theorem A.7.)

Discussion of assumption. This assumption holds for several common distributions, including all continuous distribution with a compact interval support; see Supplementary Material A.4. This also includes common truncated distributions like the truncated powerlaw and truncated normal distributions. For instance, when $\mathcal{P} = \mathrm{Unif}(0,1)$, we have $T_{\mathcal{P}}(\varepsilon(n)) = \varepsilon(n)$, and we can choose $\varepsilon(n) = \frac{\ln n}{n}$.

4 OVERVIEW OF PROOF TECHNIQUES

In this section, we overview of proofs of Theorems 3.1 and 3.2. The complete proofs have been delegated to the Supplementary Material due to space restrictions. We first need some additional notation. Let $\Phi(a)$ denote the expectation of $\mathrm{Beta}(a, b)$:

$$\Phi(a) \coloneqq \frac{a}{(a+b)}. \tag{14}$$

Note that Φ is an increasing and concave function. Let $X_{\max} \coloneqq \max_i X_i^{(t)}$ and $Y_{\max} \coloneqq \max_j Y_j^{(t)}$. Let $U_X^{(t)}$ and $U_Y^{(t)}$ be the utility of all X-candidates selected and all Y-candidates selected in iteration t respectively:

$$U_X^{(t)} \coloneqq \sum\nolimits_{i \in G_X \cap S_\ell^{(t)}} X_i^{(t)} \;\text{ and }\; U_Y^{(t)} \coloneqq \sum\nolimits_{j \in G_Y \cap S_\ell^{(t)}} Y_j^{(t)}.$$

It will be convenient to consider $\delta^{(t)} \coloneqq \frac{U^{(t)}}{\widetilde{U}^{(t)}} - 1$. Expressing the update rule (7) with $\delta^{(t)}$ we get

$$a^{(t+1)} = a^{(t)} \cdot (1 + \delta^{(t)}).$$

We can equivalently write $\delta^{(t)}$ as:

$$\delta^{(t)} = \frac{(1-\beta^{(t)}) \cdot U_X^{(t)}}{\beta^{(t)} U_X^{(t)} + U_Y^{(t)}}. \tag{15}$$

Observe that $\delta^{(t)}$ is a decreasing function of $U_Y^{(t)}$ and $\beta^{(t)}$, and an increasing function of $U_X^{(t)}$. Further, as discussed in the Section 2, $U^{(t)} \geq \widetilde{U}^{(t)}$, and so, $\delta^{(t)} \geq 0$ and $a^{(t+1)} \geq a^{(t)}$.

4.1 Proof sketch of Theorem 3.1

Fix an iteration $t \in \mathbb{N}$. In the proof, we show a lower bound on $a^{(t)}$ which holds with high probability. Then, using the fact that $\Phi(\cdot)$ is increasing, we can condition on the event that "$a^{(t)}$ is large," to derive the desired lower bound on $\mathbb{E}[\beta^{(t)}] = \mathbb{E}_{a^{(t)}}[\Phi(a^{(t)})]$.

Towards this, consider the following events:

$$\mathcal{E}^{(t)} \coloneqq \left(X_{\max}^{(t)} > Y_{\max}^{(t)}\right) \text{ and } \mathcal{F}^{(t)} \coloneqq \left(\beta^{(t)} \leq \mathrm{median}(\beta^{(t)})\right).$$

Conditioning on them we can show a one-step lower bound on $a^{(t)}$.

Lemma 4.1 (**Conditional lower bound on $a^{(t)}$**). *For all $s \in \mathbb{N}$, if $\ell > 0$, then $(\mathcal{E}^{(s)} \wedge \mathcal{F}^{(s)})$ implies*

$$\left(a^{(s+1)} > a^{(s)} + \frac{a^{(1)}(b-1)}{(k-\ell+1)(a^{(1)}+b)}\right).$$

Proof outline: First, we show that $U_Y^{(s)}/U_X^{(s)} \leq k - \ell$ conditioned on $\mathcal{E}^{(s)}$. Then, we show an upper bound on the median of the beta distribution, such that, given $\mathcal{F}^{(s)} \coloneqq (\beta^{(s)} \leq \mathrm{median}(\beta^{(s)}))$, it holds

$$\frac{(1-\beta^{(t)})}{\beta^{(t)} + (k-\ell)} \geq \frac{(b-1)}{(k-\ell+1)(a+b)}. \tag{16}$$

Now, conditioning on $\mathcal{E}^{(s)}$ and $\mathcal{F}^{(s)}$ and using the above, we get

$$\delta^{(s)} = \frac{(1-\beta^{(t)})}{\beta^{(t)} + U_Y^{(s)}/U_X^{(s)}} \geq \frac{(1-\beta^{(t)})}{\beta^{(t)} + (k-\ell)} \overset{(16)}{\geq} \frac{(b-1)}{(k-\ell+1)(a^{(t)}+b)}.$$

The lemma follows by using the update rule and simplifying.

Next, we extend Lemma 4.1 along with other facts to show a lower bound on $a^{(t)}$ (alluded to earlier). Towards this, note that $Pr[\mathcal{F}^{(t)}] = \frac{1}{2}$. We show that $\Pr[\mathcal{E}^{(t)}] = \rho$ by analyzing an equivalent urn model. Let $Z^{(t)} \in \{0, 1\}$ be the indicator random variable that $(\mathcal{E}^{(t)} \wedge \mathcal{F}^{(t)})$ occurs. Since $\mathcal{E}^{(t)}$ and $\mathcal{F}^{(t)}$ are independent,

$$\Pr[Z^{(t)} = 1] \coloneqq \Pr[\mathcal{E}^{(t)} \wedge \mathcal{F}^{(t)}] = \Pr[\mathcal{F}^{(t)}] \cdot \Pr[\mathcal{E}^{(t)}] = \frac{\rho}{2}. \tag{17}$$

Note that, the event $\mathcal{F}^{(t)}$ only depends on the draw of $\beta^{(t)}$[3] and that $\mathcal{E}^{(t)}$ only depends on $X^{(t)}$ and $Y^{(t)}$. Thus, it follows that $Z^{(t)}$ only depends on random variables from the t-th iteration, and that, for all $t_1 \neq t_2$, $Z^{(t_1)}$ and $Z^{(t_2)}$ are independent random variables. This allows us to use the Chernoff bound on the sum $\sum_{s=1}^{t} Z^{(s)}$:

$$\begin{aligned} \Pr\left[\sum\nolimits_{s=1}^{t} Z^{(s)} \leq \frac{t\rho}{4}\right] &\overset{(17)}{=} \Pr\left[\sum\nolimits_{s=1}^{t} Z^{(s)} \leq \left(1 - \frac{1}{2}\right) \cdot \mathbb{E}\left[\sum\nolimits_{s=1}^{t} Z^{(s)}\right]\right] \\ &\leq \exp\left(-\frac{1}{2^2 \cdot 2} \cdot \frac{t\rho}{2}\right). \end{aligned} \tag{18}$$

Using Lemma 4.1 and $(Z^{(s)} = 1) \coloneqq (\mathcal{E}^{(s)} \wedge \mathcal{F}^{(s)})$ for each $s \in [t]$, we get

$$a^{(t)} \geq a^{(1)} + \left(\sum\nolimits_{s=1}^{t} Z^{(s)}\right) \cdot \frac{a^{(1)}(b-1)}{(k-\ell+1)(a^{(1)}+b)}.$$

Substituting this in Equation (18) we get

$$\Pr\left[a^{(t)} \geq a^{(1)} + \frac{t\rho}{4} \frac{a^{(1)}(b-1)}{(k-\ell+1)(a^{(1)}+b)}\right] \geq 1 - \exp\left(-\frac{t\rho}{16}\right). \tag{19}$$

Let C be the constant $C \coloneqq \frac{a^{(1)}(b-1)}{4(a^{(1)}+b)} > 0$, and $\hat{a} \coloneqq a^{(1)} + C\frac{t\rho}{(k-\ell+1)}$. (Recall that $a^{(1)}$ and b are fixed constants strictly larger than 1.) Now we can prove the result as follows.

$$\begin{aligned} \mathbb{E}\left[\beta^{(t)}\right] &= \mathbb{E}_{a^{(t)}}\left[\Phi(a^{(t)})\right] \\ &= \int_{a^{(1)}}^{\hat{a}} \Phi(a)\, d\Pr[a^{(t)} = a] + \int_{\hat{a}}^{\infty} \Phi(a)\, d\Pr[a^{(t)} = a] \\ &\geq \Phi(\hat{a}) \cdot \int_{\hat{a}}^{\infty} d\Pr[a^{(t)} = a] \qquad (\Phi \text{ is increasing}) \\ &\overset{(19)}{\geq} \Phi(\hat{a}) \cdot \left(1 - \exp\left(-\frac{t\rho}{16}\right)\right) \\ &= \frac{\hat{a}}{\hat{a}+b} \cdot \left(1 - \exp\left(-\frac{t\rho}{16}\right)\right) \\ &\overset{(14)}{=} \left(\frac{a^{(1)} + C \cdot \frac{t\rho}{(k-\ell+1)}}{a^{(1)} + b + C \cdot \frac{t\rho}{(k-\ell+1)}}\right) \cdot \left(1 - \exp\left(-\frac{t\rho}{16}\right)\right) \\ &\geq \left(\frac{C \cdot \frac{t\rho}{(k-\ell+1)}}{a^{(1)} + b + C \cdot \frac{t\rho}{(k-\ell+1)}}\right) \cdot \left(1 - \exp\left(-\frac{t\rho}{16}\right)\right) \end{aligned} \tag{20}$$

Choosing $C_1 \coloneqq \frac{C}{(a^{(1)}+b)}$ and simplifying completes the proof.

Remark 4.2. *The only properties of the beta distribution used in the proof, are that $\Phi(\cdot)$ is an increasing function, and that the median of $\beta^{(t)}$ is upper bounded by a suitable value. Abstracting these properties will give an analogous proof of Equation* (10)*.*

Remark 4.3. *It might be possible to tighten the dependence of Equation* (9) *on ℓ using techniques from [14]. In particular, the factor $\frac{t\rho}{k-\ell+1}$ may improve to, roughly, $\ell \cdot \frac{t\rho}{k-\ell+1}$.*

[3] Specifically, the inverse CDF of the draw of $\beta^{(t)}$.

4.2 Proof sketch of Theorem 3.2

Fix an iteration $t \in \mathbb{N}$. In the proof, we show an upper bound on $\mathbb{E}[a^{(t)}]$. Then, using that $\Phi(\cdot)$ is concave and increasing, we can prove the desired upper bound on $\mathbb{E}[\beta^{(t)}] = \mathbb{E}_{a^{(t)}}[\Phi(a^{(t)})]$.

Let $D^{(t)}$ denote the set of all random variables from the first $t-1$ iterations:

$$D^{(t)} := \left\{\beta^{(s)}\right\}_{1\leq s<t} \cup \left\{X^{(s)}\right\}_{1\leq s<t} \cup \left\{Y^{(s)}\right\}_{1\leq s<t}.$$

Then, we prove the following lemmas.

Lemma 4.4 (**Upper bound on $\delta^{(t)}$ when $\beta^{(t)}$ is close to 1**). *For all iterations $s \in \mathbb{N}$, constant $\varepsilon \in (0, 1/2)$, and values of $D^{(s)}$, if $\beta^{(s)} \geq 1-\varepsilon$, then $\delta^{(s)} \leq 2\varepsilon$.*

Proof: $\delta^{(s)}$ is an increasing function of $U_Y^{(s)}/U_X^{(s)}$. In the worst case we have $U_Y^{(s)}/U_X^{(s)} = 0$. Substituting this in Equation (15):

$$\delta^{(s)} \overset{(15)}{=} \frac{(1-\beta^{(s)})\cdot U_X^{(s)}}{\beta^{(s)}U_X^{(s)} + U_Y^{(s)}} \leq \frac{(1-\beta^{(s)})}{\beta^{(s)}} \leq \frac{\varepsilon}{1-\varepsilon} \overset{(\varepsilon<\frac{1}{2})}{\leq} 2\varepsilon.$$

Lemma 4.5 (**Upper bound on $\mathbb{E}\left[\delta^{(t)}\right]$ when $\beta^{(t)}$ is not close to 1**). *For all iterations $s \in \mathbb{N}$, constant $\varepsilon \in (0, 1/2)$, values of $D^{(s)}$, and parameters $a^{(1)}, b > 1$, it holds*

$$\Pr[\beta^{(s)} < 1-\varepsilon] \cdot \mathbb{E}\left[\delta^{(s)} \,\middle|\, (\beta^{(s)} < 1-\varepsilon) \wedge (U_X^{(t)} \neq 0)\right] \leq \frac{b}{a^{(1)}-1}.$$

Due to space constraints we defer the proof of Lemma 4.5 till Supplementary Material A.3.1. It roughly follows by upper bounding $\delta^{(s)}$ by $(\beta^{(s)})^{-1}$, rearranging, and using the value of the inverse moment of the beta distribution.

Lemma 4.6 (**Upper bound on $\Pr\left[U_X^{(t)} \neq 0\right]$ when $\beta^{(t)}$ is not close to 1**). *For all iterations $t \in \mathbb{N}$ and values of $D^{(s)}$, if $\ell = 0$, then there exists an $n_0 \in \mathbb{N}$, such that, for all $n \geq n_0$ and $\varepsilon \leq \frac{\ln n}{n}$, it holds*

$$\Pr\left[U_X^{(t)} \neq 0 \,\middle|\, \beta^{(t)} < 1-\varepsilon\right] \leq \exp\left(-\frac{1}{8}\varepsilon n(1-\rho)\right).$$

Proof outline: In the proof, we show that the decision-maker selects an X-candidate with probability at most $\exp\left(-\frac{1}{8}\varepsilon n(1-\rho)\right)$. Notice that this implies the lemma, since if no X-candidates are selected then $U_X^{(s)} = 0$. Towards this, we use the Chernoff bound to show that, conditioned on $\beta^{(s)} < 1-\varepsilon$, with high probability at least k Y-candidates have higher observed utilities than $X_{\max}$.
(Note that Lemma 4.6 uses $\ell = 0$.) We would like to upper bound $\mathbb{E}\left[\delta^{(s)} \,\middle|\, D^{(s)}\right]$ for all values of $D^{(s)}$. To this end, consider three cases:
(Case 1) $\beta^{(s)}$ is close to 1 ($\beta^{(s)} \geq 1-\varepsilon$),
(Case 2) $\beta^{(s)}$ is far from 1 ($\beta^{(s)} < 1-\varepsilon$) and $U_X^{(t)} \neq 0$, and
(Case 3) $\beta^{(s)}$ is far from 1 ($\beta^{(s)} < 1-\varepsilon$) and $U_X^{(t)} = 0$.
Using Lemma 4.4, in Case 1 we have $\delta^{(s)} \leq 2\varepsilon$. In Case 2, from Lemma 4.5 and Lemma 4.6, we can bound $\mathbb{E}\left[\delta^{(s)} \,\middle|\, D^{(s)}\right]$ by $\frac{b}{a^{(1)}-1} \cdot \exp\left(-\frac{1}{8}\varepsilon n(1-\rho)\right)$. Finally, in Case 3 it holds that $\delta^{(s)} = 0$ (see Equation (15)). Note that, each of these results holds for all values of $D^{(s)}$. Combining these and setting $\varepsilon := \frac{8\ln n}{n(1-\rho)}$ we can show that

$$\mathbb{E}\left[\delta^{(s)} \,\middle|\, D^{(s)}\right] \leq \frac{16\ln n}{n(1-\rho)} \cdot \frac{a^{(1)}+b}{a^{(1)}-1}. \tag{21}$$

Using same $\mathbb{E}\left[\delta^{(s)}\right] = \mathbb{E}\left[\mathbb{E}\left[\delta^{(s)} \,\middle|\, D^{(s)}\right]\right]$, upper bound also holds for $\mathbb{E}\left[\delta^{(s)}\right]$. Consider $t_1 < t_2$. Using Equation (21) we have

$$\begin{aligned}
\mathbb{E}\left[\delta^{t_1}\cdot\delta^{t_2}\right] &= \int_0^\infty x \cdot \mathbb{E}\left[\delta^{t_2} \mid \delta^{t_1} = x\right] \cdot \Pr[\delta^{t_1} = x]dx \\
&\overset{(21)}{\leq} \left(\frac{16\ln n}{n(1-\rho)} \cdot \frac{a^{(1)}+b}{a^{(1)}-1}\right) \cdot \int_0^\infty x \cdot \Pr[\delta^{t_1} = x]dx \\
&\overset{(21)}{\leq} \left(\frac{16\ln n}{n(1-\rho)} \cdot \frac{a^{(1)}+b}{a^{(1)}-1}\right)^2.
\end{aligned}$$

Similarly for $t_1 < t_2 < \cdots < t_s$ we can show that

$$\mathbb{E}\left[\prod\nolimits_{i\in[s]}\left(1+\delta^{t_i}\right)\right] \leq \left(1 + \frac{16\ln n}{n(1-\rho)} \cdot \frac{a^{(1)}+b}{a^{(1)}-1}\right)^s. \tag{22}$$

Pick an n_0, such that, $\frac{16\ln n}{n(1-\rho)} \cdot \frac{a^{(1)}+b}{a^{(1)}-1} \leq \frac{1}{t}$. Then, we can show:

$$\mathbb{E}\left[a^{(t)}\right] = a^{(1)}\mathbb{E}\left[\prod_{i=1}^{t-1}\left(1+\delta^{t_i}\right)\right] \leq a^{(1)} + 2t\frac{16\ln n}{n(1-\rho)} \cdot \frac{a^{(1)}(a^{(1)}+b)}{a^{(1)}-1}. \tag{23}$$

Further, we have

$$\begin{aligned}
\mathbb{E}[\beta^{(t)}] = \mathbb{E}_{a^{(t)}}[\Phi(a^{(t)})] &\leq \Phi\left(\mathbb{E}_{a^{(t)}}[a^{(t)}]\right) && (\Phi \text{ is concave}) \\
&= \frac{\mathbb{E}_{a^{(t)}}[a^{(t)}]}{\mathbb{E}_{a^{(t)}}[a^{(t)}]+b}. && (\Phi \text{ is increasing})
\end{aligned}$$

Now substituting Equation (23) in the above, rearranging, and choosing $C_2 := \frac{32a^{(1)}}{a^{(1)}-1}$, we can complete the proof.

Remark 4.7. *The only properties of the beta distribution used in the proof, are that $\Phi(\cdot)$ is concave and increasing and an upper bound on the inverse moment of the beta distribution. Abstracting these properties gives an analogous proof for Equation* (11).

5 EMPIRICAL OBSERVATIONS

We enlisted participants on Amazon Mechanical Turk for an iterative candidate selection experiment in order to observe the effect of the Rooney Rule on their long-term behavior. Participants were shown colored tiles representing candidates, along with the tiles' *observed* utilities, and were instructed to select a small fixed number of tiles with the goal of maximizing the *latent* utility of the selection; this was incentivized by tying their bonus payment to the latent utility attained. They were informed that the observed utilities were noisy estimates of the latent utilities. We required some participants to follow the Rooney Rule and did not constrain others. In this section, we give an overview of the experimental design and our analysis of the results.

5.1 Experimental design

A total of 76 participants located in the U.S. were recruited using Amazon Mechanical Turk to complete a repeated "selection" experiment.[4] In each of $T = 25$ iterations, participants were presented with $n = 100$ tiles, representing candidates, in order of highest to lowest observed utility. Half of the tiles were red (representing Y-candidates) and half were blue (representing X-candidates); i.e., $\rho = \frac{1}{2}$. The latent utilities, $X_i^{(t)}, Y_j^{(t)}$ for each tile was drawn from Unif$(0, 100)$, the uniform distribution on the interval $[0, 100]$. The observed utilities for the red tiles $\widetilde{Y}_j^{(t)}$ were sampled from $\mathcal{N}(Y_j^{(t)}, 3)$.[5] The observed utilities for the blue tiles ($\widetilde{X}_i^{(t)}$) were sampled from $\mathcal{N}(\beta \cdot X_i^{(t)}, 3)$ where $\beta = \frac{2}{3}$ is the bias coefficient.[6] All observed and latent utilities were then rounded to the nearest integer.

[4] Participants were paid at a rate of about $15/hour. The experiment received institutional review board (IRB) exemption, Category 2 and 3, from Yale University.

The participants were instructed to select $k = 10$ tiles to try to maximize the total latent utility $U^{(t)}$ in each iteration $t \in [T]$, given the observed utility of each tile. They were told that the observed utilities were noisy estimates of the latent utilities. In each round, after they submitted their choices, participants were shown the latent utility of each tile they selected; hence, they could learn about the latent utilities of each group over time.[7]

About half of the participants (39 out of 76), chosen at random, were constrained to follow the Rooney Rule with $\ell = 1$, i.e., they were required to select at least one blue tile. The other half of the participants (37 out of 76) were unconstrained (i.e., $\ell = 0$). We refer to the constrained group of participants as RRGrp and the unconstrained group of participants as ControlGrp. Our goal was to compare the performance across the two groups RRGrp and ControlGrp in the long term, i.e., after having been granted some opportunity to learn about the latent utilities of both groups. Hence, our analysis focuses on the choices made in the last fifteen iterations.[9]

5.2 The Rooney Rule results in less inequality in the long term

Recall that participants selected $k = 10$ tiles in each iteration and that, according to the distribution of latent utilities, participants should select five blue and five red tiles on average, the observed utilities exhibited bias against the blue tiles, rendering them lower in the ranking and hence less likely to be selected. In the last fifteen observations, we observe that the RRGrp selects 4.0 blue tiles on average, while ControlGrp selects 2.5 on average.

As participants subject to the Rooney Rule are required to select at least one blue tile, one might wonder if this difference between the RRGrp and ControlGrp is solely due to the Rooney Rule requirement. Surprisingly, this is not the case, suggesting that additional learning has occurred. In particular, we consider the number of blue tiles selected *in addition to* the ones required by the Rooney Rule; i.e., we subtract 1 from the number of blue tiles selected for RRGrp participants and 0 from ControlGrp. We find a statistically significant ($p < 0.001$) difference between the number of blue tiles selected in addition to the requirement between RRGrp and ControlGrp (see Figure 1).

This suggests that participants "learn" to select more equal numbers of blue and red tiles as a result of implementing the Rooney Rule. This aligns with our theoretical result that participants subject to the Rooney Rule reduce their biases at a faster rate than an unconstrained decision-maker.

5.3 The Rooney Rule does not negatively affect the utility in the long term

We observe that the total utility $U^{(t)}$ modestly increases over iterations $t \in [T]$ for both RRGrp and ControlGrp. This suggests that participants in both RRGrp and ControlGrp are learning to select more optimally over time.

However, we are particularly interested in understanding how well a decision-maker subject to the Rooney Rule can fulfill its (latent) utility-maximizing objective in the long term. To this end, we compare the mean latent utility attained by RRGrp and ControlGrp during the last fifteen iterations. We measure the latent utility achieved by RRGrp and ControlGrp as a proportion of an "optimal" latent utility that could be achieved if one knew the distributions of latent and observed utilities a priori.

Figure 1: The number of blue tiles selected in addition to the required number is greater for RRGrp than ControlGrp with significance $p < 0.001$ using Welch's t-test.[10]

	Mean	Standard deviation
RRGrp	3.0	2.4
ControlGrp	2.5	3.1

	Welch's t-test
Null hypothesis	population means same
Alternative hypothesis	population means different
t-statistic	8.2
Degrees of freedom	1800
p-value	<0.001

In a given iteration $t \in [T]$, define the *optimal strategy set* as the tiles with the top k values from $\{\widetilde{X}_i^{(t)}/\beta\}_{i \in G_X} \cup \{\widetilde{Y}_j^{(t)}\}_{j \in G_Y}$. We call the latent utility of the optimal strategy set the *optimal strategy utility*.[11] From this point on, we report the latent utility for a given selection as a fraction of the optimal strategy utility. We also define the *latent utility derived from blue (resp. red) tiles* for a given selection to mean the latent utility of selected blue (resp. red) tiles as a fraction of the optimal strategy utility.

Figure 2: The latent utility of RRGrp is 2% less than that of ControlGrp, but the latent utility derived from the blue tiles was higher for RRGrp than ControlGrp.

Latent utility: mean *(standard deviation)*

	overall	derived from blue	derived from red
RRGrp	0.87 *(0.11)*	0.37 *(0.22)*	0.50 *(0.24)*
ControlGrp	0.89 *(0.12)*	0.22 *(0.28)*	0.67 *(0.31)*

We find that, in the long term, the latent utility obtained by RRGrp is less than that of the ControlGrp by only 2%. Moreover, the latent utility derived from the blue tiles was higher for the RRGrp than ControlGrp (see Figure 2). Hence, even though the latent utility attained by RRGrp and ControlGrp were not substantially different, a much greater proportion of RRGrp latent utility came from blue tiles. (The difference between the means of latent utility derived from blue tiles for the two groups is statistically significant at $p < 0.001$. See Supplementary Material C.3 for details.) This suggests that implementing the Rooney Rule can increase the contributions from underrepresented candidates to the overall latent utility without substantially decreasing the total latent utility attained.

As another metric for how well a decision-maker subject to the Rooney Rule can fulfill its utility-maximizing objective, we also analyzed *optimal strategy overlap*, the number of selected tiles in the optimal strategy set. We find similar results: the total optimal strategy overlap is not very different for RRGrp versus ControlGrp, but the optimal strategy overlap derived from to blue tiles is much higher for RRGrp (see Supplementary Material C.4 for details).

[6] $\mathcal{N}(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ.

[7] We choose $\beta = 2/3$ because previous empirical work has found evidence for bias at least as extreme as 2/3 [65].

[8] For images of the user interface, the instructions, the demonstration, and feedback, we refer the reader to Supplementary Material C.1.

[9] This number is arbitrary to some extent, but we found similar results for other choices.

[10] Welch's t-test assumes independence between samples. We also analyzed mixed effects model to account for the multiple samples from each participant and found similar results (see Supplementary Material C.2).

[11] Because of the noise added to each observed utility, the optimal strategy may not be the same as the set of tiles with the greatest latent utilities.

5.4 Summary of empirical results

We observe two dominant trends in participant behavior in the empirical iterative selection experiment:

- Participants subject to the Rooney Rule select more blue tiles *in addition to* the required number, in the long term, than the unconstrained participants. We interpret this to mean that they learn to select more equal numbers of blue and red tiles faster than the unconstrained participants. This aligns with our main theoretical result that participants subject to the Rooney Rule reduce their biases at a faster rate than others.
- There was no substantial difference in the mean latent utility of RRGrp versus ControlGrp. However, while the mean total latent utility between RRGrp and ControlGrp was not very different, a much larger proportion of the latent utility came from blue tiles for participants subject to the Rooney Rule. We interpret this to mean that participants subject to the Rooney Rule learn to increase the latent utility derived from blue tiles while nearly matching the total latent utility of selections made by unconstrained participants.

These trends suggest that, over the long term, implementing the Rooney Rule can yield significant benefits for increasing representation of underrepresented candidates without substantially decreasing total latent utility.

6 LIMITATIONS

Model and theoretical results. We consider a model with two groups, where the decision-maker is implicitly biased against one underrepresented group. However, in real-world applications, there may be multiple and intersecting groups (say, those defined by race and gender) and where the candidates at the intersection of two or more underrepresented groups may face a larger implicit bias [7, 66].

Further, we consider that the decision-maker evaluates the candidates selected (say, for an interview) in aggregate (it compares $U^{(t)}$ and $\widetilde{U}^{(t)}$) and sees the latent utilities of all selected candidates without noise. But, the decision-maker may assign more or less weight to the latent utility of the X-candidates and learn only a noisy version of the latent utility of the selected candidates. We make some progress toward this by allowing for noise in the decision-maker's implicit bias (see Section 3.2.1).

Empirical results. The iterative selection experiment in our empirical results is different from real-world candidate selection tasks in several ways: For one, the bias β is imposed exogenously on the observed utilities that the selection decision-maker sees, rather than coming from their own implicit cognitive processes. If selection decision-makers evaluate the merit of candidates based on their own experience or criteria, they might trust these assessments more than the observed utilities presented to them in the experiment. Similarly, the stakes of the experiment for participants were lower than in real-world tasks; decision-makers in the real world may pay more attention to slight variations in the performance of the candidates selected.

In future work, it would be interesting to study how participant biases change over time if they assess candidates based on subjective criteria (like a recommendation letter) that is biased against one group, rather than being provided biased observed utilities at the beginning of each iteration. It would also be valuable to incorporate the real-world social biases of participants into an iterative candidate selection task—this could shed insight on how actual social biases shape people's decisions.

Specificity. Our theoretical analysis and empirical observations are only applicable when the decision-maker is not explicitly biased and the selection process is not tokenistic. An explicitly biased decision-maker would remain biased no matter which candidates are presented, and a tokenistic process may not require the decision-maker to consider underrepresented candidates seriously.

Further, other societal and structural biases and discrimination can also influence the decision-maker, as when standardized test scores disadvantage minority and low-income candidates [27]. The Rooney Rule is only one of the many policies to mitigate bias and discrimination more broadly, and it is important to consider implementing it as a part of a larger toolkit for positive social change.

7 CONCLUSION

We consider a theoretical model of how a decision-maker's implicit bias changes over repeated candidate selection processes and study the effect of the Rooney Rule on the decision-maker's implicit bias. We show that, under our model, if the decision-maker uses the Rooney rule, then the rate at which its implicit bias reduces decreases with the size of the shortlist but is independent of the number of candidates. However, when the decision-maker does not use the Rooney Rule, the same rate decreases with the number of candidates (Section 3). Thus, in the regime where the total number of candidates is much larger than the size of the shortlist, our results predict a significantly faster reduction in the decision-maker's implicit bias by using the Rooney Rule—giving another reason to use it to mitigate implicit bias in the long term.

Toward understanding the robustness of this result, we consider some extensions of the model and show how our results generalize to them (Section 3.2). Here, in particular, we consider noise in the implicit bias of the decision-maker, which can change the distribution of the implicit bias, and identify properties of the resulting distribution where our results hold (see Section 3.2.1). An interesting direction for future work could be to examine the effects of noise in the observed utilities—as in our empirical study (Section 5).

Our empirical findings based on the experiment on Amazon Mechanical Turk show that, over multiple rounds, the Rooney Rule helped the participants learn to select a shortlist which is more representative of the qualified candidates in the applicant pool, without substantially decreasing the latent utility of selected candidates.

Our findings provide evidence that the Rooney Rule can reduce the biases of the decision-maker in repeated candidate selection processes. The policy can be a simple and effective tool as part of a larger effort to mitigate implicit bias.

ACKNOWLEDGEMENTS

This research was supported in part by a J.P. Morgan Faculty Award and an AWS MLRA grant.

REFERENCES

[1] Daron Acemoglu, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

[2] Daron Acemoglu and Asuman Ozdaglar. Opinion dynamics and learning in social networks. *Dynamic Games and Applications*, 1(1):3–49, 2011.

[3] Elizabeth Anderson. *The imperative of integration*. Princeton University Press, Princeton, 2013.

[4] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012.

[5] N. Balakrishnan, N.L. Johnson, and S. Kotz. *Continuous Univariate Distributions*. Wiley Series in Probability and Statistics. John Wiley & Sons Incorporated, Hoboken, NJ, 2016.

[6] Mark W. Bennet. Unraveling the gordian knot of implicit bias in jury selection: The problems of judge-dominated voir dire, the failed promise of batson, and proposed solutions. *Harvard Law and Policy Review*, 4(1):149–172, 2010.

[7] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.

[8] Irene V. Blair, Jennifer E. Ma, and Alison P. Lenton. Imagining stereotypes away: the moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology*, 81(5):828, 2001.

[9] Scott Bland. Schumer to introduce rules for diverse senate hiring. *Politico*, 2017. https://www.politico.com/story/2017/02/schumer-diversity-nfl-rooney-rule-235477.

[10] Jerome R. Busemeyer, Eunhee Byun, Edward L. Delosh, and Mark A. McDaniel. Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In Koen Lamberts and David Shanks, editors, *Knowledge, Concepts, and Categories*, pages 405–437. MIT Press, Cambridge, MA, 1997.

[11] L. Elisa Celis, Chris Hays, Anay Mehrotra, and Nisheeth K. Vishnoi. Code for Amazon Mechanical Turk experiment for "The Effect of the Rooney Rule on Implicit Bias in the Long Term", October 2020. https://github.com/johnchrishays/downstream-rooney-mturk.

[12] L. Elisa Celis, Chris Hays, Anay Mehrotra, and Nisheeth K. Vishnoi. Reward maximization game, October 2020. https://downstreamrooney.herokuapp.com/experiment/.

[13] L. Elisa Celis, Peter M. Krafft, and Nisheeth K. Vishnoi. A distributed learning dynamics in social groups. In Elad Michael Schiller and Alexander A. Schwarzmann, editors, *Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC 2017)*, pages 441–450. ACM, 2017.

[14] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. Interventions for ranking in the presence of implicit bias. In *Proceedings of the ACM Conference on Fairness, Accountability and Transparency (FAT* 2020)*, pages 369–380. ACM, 2020.

[15] Elizabeth N. Chapman, Anna Kaatz, and Molly Carnes. Physicians and implicit bias: How doctors may unwittinglyperpetuate health care disparities. *Journal of General Internal Medicine*, 28(11):1504–10, 2013.

[16] Tessa E.S. Charlesworth and Mahzarin R. Banaji. Patterns of implicit and explicit attitudes: I. long-term change and stability from 2007 to 2016. *Psychological Science*, 30(2):174–192, 2019.

[17] Erick Chastain, Adi Livnat, Christos Papadimitriou, and Umesh Vazirani. Algorithms, games, and evolution. *Proceedings of the National Academy of Sciences*, 111(29):10620–10623, 2014.

[18] Bernard Chazelle and Chu Wang. Iterated learning in dynamic social networks. *The Journal of Machine Learning Research*, 20(1):979–1006, 2019.

[19] Brian W. Collins. Tackling unconscious bias in hiring practices: The plight of the Rooney Rule. *New York University Law Review*, 82(3):870–912, 2007.

[20] Alexander Coutts. Good news and bad news are still news: experimental evidence on belief updating. *Experimental Economics*, 22(2):369–395, 2019.

[21] Nilanjana Dasgupta. Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. In *Advances in Experimental Social Psychology*, volume 47, pages 233–279. Elsevier, Amsterdam, 2013.

[22] Nilanjana Dasgupta and Anthony G. Greenwald. On the malleability of automatic attitudes: combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology*, 81(5):800, 2001.

[23] Nilanjana Dasgupta and Luis M. Rivera. 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):112–123, 2008.

[24] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[25] Megan Rose Dickey. Lyft's diversity efforts are not going unnoticed. *TechCrunch*, 2017. https://techcrunch.com/2017/09/27/lyfts-diversity-efforts-are-not-going-unnoticed/.

[26] N. Jeremi Duru. The Rooney Rule's reach: How the NFL's equal opportunity initiative for coaches inspired local government reform. *Oxford Handbook of American Sports Law*, 2018.

[27] Kim Elsesser. Lawsuit claims sat and act are biased—here's what research says. *Forbes*, 2019. https://www.forbes.com/sites/kimelsesser/2019/12/11/lawsuit-claims-sat-and-act-are-biased-heres-what-research-says/#7175371e3c42.

[28] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit variance. In *Proceedings of the 21st ACM Conference on Economics and Computation (EC 2020)*, pages 649–675. ACM, 2020.

[29] Alexander R. Green, Dana R. Carney, Daniel J. Pallin, Long H. Ngo, Kristal L. Raymond, Lisa I. Iezzoni, and Mahzarin R. Banaji. Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of General Internal Medicine*, 22(9):1231–1238, 2007.

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

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

[32] Alison V. Hall, Erika V. Hall, and Jamie L. Perry. Black and blue: Exploring racial bias and law enforcement in the killings of unarmed black male civilians. *UCLA Law Review*, 71(3):175–186, 2016.

[33] Charles A. Holt and Angela M. Smith. An update on bayesian updating. *Journal of Economic Behavior & Organization*, 69(2):125–134, 2009.

[34] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference (WWW 2018)*, pages 1389–1398. ACM, 2018.

[35] Quinn Capers IV, Daniel Clinchot, Leon McDougle, and Anthony G. Greenwald. Implicit racial bias in medical school admissions. *Academic Medicine*, 92(3):365–369, 2017.

[36] Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012.

[37] Nathalia Jimenez, Kristy Seidel, Lynn D. Martin, Frederick P. Rivara, and Anne M. Lynn. Perioperative analgesic treatment in latino and non-latino pediatric patients. *Journal of Health Care for the Poor and Underserved*, 21(1):229–236, 2010.

[38] A. Jøsang. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2016.

[39] Audun Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(3):279–311, 2001.

[40] Jerry Kang, Judge Mark Bennett, Devon Carbado, Pam Casey, Nilanjana Dasgupta, David Faigman, Rachel Godsil, Anthony G. Greenwald, Justin Levinson, and Jennifer Mnookin. Implicit bias in the courtroom. *UCLA Law Review*, 59(5):1124–1187, 2012.

[41] Mary E. Kite and Bernard E. Whitley Jr. *Psychology of prejudice and discrimination*. Psychology Press, 2016.

[42] Jon Kleinberg and Manish Raghavan. Selection problems in the presence of implicit bias. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. ACM, 2018.

[43] Kyunghee Koh and David E. Meyer. Function learning: Induction of continuous stimulus–response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*, 17(5):811–836, 1991.

[44] Ulrich Krause. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in Difference Equations*, 2000:227–236, 2000.

[45] Björn Lindström, Ida Selbing, Tanaz Molapour, and Andreas Olsson. Racial bias shapes social reinforcement learning. *Psychological Science*, 25(3):711–719, 2014.

[46] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158. PMLR, 2018.

[47] Christopher G. Lucas, Thomas L. Griffiths, Joseph J. Williams, and Michael L. Kalish. A rational model of function learning. *Psychonomic Bulletin & Review*, 22(5):1193–1215, 2015.

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

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

[50] Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge university press, 1995.

[51] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the ACM Conference on Fairness, Accountability and Transparency (FAT* 2019)*, pages 359–368. ACM, 2019.

[52] Christina Passariello. Tech firms borrow football play to increase hiring of women. *Wall Street Journal*, 2016. https://www.wsj.com/articles/tech-firms-borrow-football-play-to-increase-hiring-of-women-1474963562.

[53] B. Keith Payne, Heidi A. Vuletich, and Jazmin L. Brown-Iannuzzi. Historical roots of implicit bias in slavery. *Proceedings of the National Academy of Sciences*, 116(24):11693–11698, 2019.

[54] Mark E. Payton, Linda J. Young, and J.H. Young. Bounds for the difference between median and mean of beta and negative binomial distributions. *Metrika*, 36(1):347–354, 1989.

[55] Julie R. Posselt. *Inside graduate admissions: Merit, diversity, and faculty gatekeeping.* Harvard University Press, Cambridge, MA, 2016.

[56] M. Amin Rahimian and Ali Jadbabaie. Learning without recall from actions of neighbors. In *American Control Conference (ACC)*, pages 1060–1065. IEEE, 2016.

[57] Jason Reid. Rethinking the NFL's Rooney Rule for more diversity at the top. *Five Thirty Eight*, 2016. https://fivethirtyeight.com/features/rethinking-the-nfls-rooney-rule-for-more-diversity-at-the-top/.

[58] Dan-Olof Rooth. Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, 17(3):523–534, 2010.

[59] Laurie A. Rudman. Sources of implicit attitudes. *Current Directions in Psychological Science*, 13(2):79–82, 2004.

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

[61] Howard Schuman, Charlotte Steeh, Lawrence Bobo, and Maria Krysan. *Racial attitudes in America: Trends and interpretations.* Harvard University Press, Cambridge, 1997.

[62] Boris Škorić, Sebastiaan JA de Hoogh, and Nicola Zannone. Flow-based reputation with uncertainty: evidence-based subjective logic. *International Journal of Information Security*, 15(4):381–402, 2016.

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

[64] David Waldstein. Success and shortfalls in effort to diversify n.f.l. coaching. *New York Times*, 2015. https://www.nytimes.com/2015/01/21/sports/football/jets-hiring-of-todd-bowles-leaves-nfl-far-short-of-goal-on-diversity.html.

[65] Christine Wenneras and Agnes Wold. Nepotism and sexism in peer-review. *Nature*, 387:341–343, 1997.

[66] Joan C. Williams. Double jeopardy? an empirical study with implications for the debates over implicit bias and intersectionality. *Harvard Journal of Law & Gender*, 37:185, 2014.

[67] Jabari Young. NFL wants to combat diversity problems with data system in Rooney Rule expansion. *CNBC*, May 2020. https://www.cnbc.com/2020/05/24/nfl-wants-to-combat-diversity-problems-with-data-system-in-rooney-rule-expansion.html.

[68] Jonathan C. Ziegert and Paul J. Hanges. Employment discrimination: The role of implicit attitudes, motivation, and a climate for racial bias. *Journal of Applied Psychology*, 90(3):553–562, 2005.