Implicit Bias in Dynamic Selection Problems

Dean Li
Advisor: Yang Cai
May 2, 2019

Abstract. Experimental studies have demonstrated that implicit bias is an important driver of discriminatory and unfair outcomes in a wide range of selection-based activities: from hiring decisions and school admissions to voting and prediction of recidivism. Beyond the effects of implicit bias on on fairness, in this paper we consider the inefficiencies that implicit bias introduces in selection problems and how we might analyze these without exact knowledge of the form of the bias or certainty that the bias exists. We extend the multiple-choice secretary problem to create a theoretical framework for studying the effects of implicit bias on dynamic selection problems. Using a general form of bias, we demonstrate the significant costs of even “low” levels of implicit bias on success rates of a class of dynamic selection algorithms. We then propose potential affirmative action selection mechanisms for remedying these costs, and we characterize their optimal forms. Finally, we demonstrate that proposing, optimizing, and simulating the utility of these mechanisms is computationally simple even over complex beliefs, and we characterize scenarios in which these affirmative actions mechanisms increase expected utility.

The goal of this work is to improve our understanding of the effects of implicit bias on group-blind on a class of dynamic selection algorithms. In addition, we characterize a framework through which we can examine how diversity policies might improve not only expected fairness but expected utility. In particular, this work serves to nuance and extend the conversation about the role of protected classes and affirmative action measures in algorithms. A large body of legal literature has been written on anti-classification, where protected classes are not used to make decisions. This can be extended to orthogonalizing other predictors before dropping out the protected classes. However, in our work, we show that group-aware mechanisms can improve both expected utility and classification parity.

1 Introduction

1.1 Motivation

In the last fifty years, the Supreme Court has weighed in on educational affirmative action policies five times with a sixth appearance looming.[1] This is no accident: from well-documented diversity issues at Stuyvescent[2] to controversies over recidivism prediction algorithms, fairness and how we deal with groups that might face bias and discrimination have been an increasingly important areas of research.

Traditional literature has often cast fairness as a constraint under which we must optimize with the implication that the “cost” of fairness is the distance between the unconstrained and constrained maximization problem.[3] Recent literature has focused extensively on analyzing the optimal trade-off in practice and theory.[4][5] Under these frameworks, we can interpret affirmative action policies – which may call for separate treatment or thresholds for a protected group – as utility-decreasing measures which enable us to satisfy a requirement for certain measures of fairness.

Other literature has studied the positive effects of diversity and how we might optimize with such preferences. For example, it demonstrable that with a utility function that is additive in some measure of predictive accuracy and diversity, the use of protected group status in algorithms must strictly improve utility.[6] The effects of diversity in settings such as the workplace have been extensively studied, although empirical results has not always been conclusive in differentiating the positive and negative effects.[7] In this context, we might understand affirmative action policies as tools that allow us to maximize a preference for diversity.
In this paper, however, I aim to add to the growing body of literature that interprets affirmative action as a measure to counter implicit bias and cover some of its cost on expected utility. This argument allows diversity policies to be more palatable to social planners who can acknowledge implicit bias but believe that diversity policies are an unacceptable violation of merit.[8] In my model, I demonstrate the effect of implicit bias on the success rate of a simple class of algorithms and characterize the cases in which simple affirmative action mechanisms increase expected utility. Finally, we demonstrate the simplicity of calculating the expected value of our affirmative action policies over arbitrary beliefs defined over biases and whether they exist.

This builds on previous work which examines one such affirmative action policy in a static context and assumes a Pareto distribution for the quality of candidates.[9] Rather than the static problem, I examine a dynamic problem with only ordinal rankings and no prior distribution over the candidates. In addition, I extend the scenario to where implicit bias may or may not exist and I demonstrate that our form of bias closely approximates the impact of any form of negative bias against minority candidates that preserves the relative rankings.

1.2 Implicit Bias

One of the keys to the model is the possible existence of implicit bias. Thus, we provide a thorough intuitive definition before we formalize implicit bias in Section 2. We propose a common definition of implicit bias as discrimination outside of the evaluating agent’s awareness.[10] However, we emphasize the fact that the agent does not have to apply discrimination – implicit bias functions as any factors outside of the knowledge of the agent which might influence the agent’s evaluation of the ranking of the candidate.

For example, consider a college admissions officer reading two different applications. One of these students might have been valedictorian, while another one of the students may have been in the 25th percentile. The admissions officer may interpret the former as far more impressive than the later without taking into account the competitiveness of the high schools. Other examples may be more subtle: consider two recommendations from the same professor. If the professor were biased against one of the students based on a characteristic not relevant to the quality of the student, this information would not be accessible to the admissions officer. If she made a decision based on the recommendation, we would characterize the gap between the true and perceived quality of the discriminated student as implicit bias.

Research has demonstrated that implicit bias has measurable effects in fields from employment and education to law and even politics.[11] In addition, optimizing in the presence of implicit bias is complex due to its nature. Because it is outside of the knowledge of the evaluating agent, its existence and severity must be assigned probability distributions rather than numerical values.

1.3 The Secretary Problem

Because we aim to study the effects of implicit bias in dynamic settings, we begin with one of the most established dynamic selection problems: The Secretary Problem, also known as the marriage problem, the Googol game, and the best choice problem concerns the optimal strategy in determining the correct time to take a particular action, in order to maximize an expected reward. This problem forms the basis of our model, and a description of the game is given below.

A manager wants to hire the best candidate out of n ordinally rankable applicants for a position. She interviews them in random order, one at a time. After interviewing a candidate, she can compare them to all of the candidates that she has already seen, but has no additional information about the distribution of the remaining candidates. She must make a decision about each applicant immediately after the interview. After a decision is made, it is irreversible. The process ends after a candidate is hired or all n candidates have been interviewed. If the best candidate has been hired, the process is successful. Otherwise, it is a failure.
The optimal strategy has a particularly elegant solution. The manager interviews \( \frac{n}{2} \) candidates in an “exploration phase,” and these candidates are rejected immediately. Naturally, if the best candidate was among the initial pool, the process ends in failure. In the next phase, a “selection phase”, the manager will hire the next candidate who ranks higher than all of the candidates from the exploration phase. This strategy guarantees the optimal chance of success: \( \frac{1}{n} \).

The history of this problem and its many formulations has been elegantly covered. Researchers have continued to study extensions of the problems, from cases where the size of the pool of candidates is unknown to cases in which the goal is to hire the second-best candidate.

The “multi-choice” extension of this problem represents a scenario where the manager can hire \( k \) candidates from the pool. There are several objective functions that can be used to evaluate the success of this process. For example, the process may be deemed successful only if the \( k \) candidates hired are the \( k \) best candidates. We will examine the extension such that the process is deemed successful if the best candidate in contained in the set of \( k \) candidates chosen. This can be viewed as the dynamic version of choosing the \( k \) short-listed candidates for a job: each candidate, when they are interviewed, must be notified within a certain amount of time and the manager cannot go back and change the decision after the interview.

2 Theoretical Framework

2.1 The Model

Suppose we have \( n \) candidates to evaluate. This set of candidates is ordinally ranked \( \{1, \ldots, n\} \) such that the best candidate is ranked 1. The arrival order of the candidates is drawn from a uniform distribution over the set of possible permutations of \( \{1, \ldots, n\} \), and we would like to hire \( k \) candidates such that \( k < n \). We must make a decision on each candidate as soon as they arrive, and decisions are irreversible.

We define the four major aspects of the model below: the utility function, the class of algorithms that we will examine, the group labels, and the bias.

Utility Function.

We measure success in terms of whether or not the best candidate is contained within the set of our hired candidates. Our algorithm is successful if the best candidate is selected and unsuccessful otherwise. We study the asymptotic success rates of algorithms – namely, the success rate as \( n \to \infty \).

The Class of Algorithms.

We examine the class of candidate-counting algorithms that are both static and simple. Previous work on the multi-choice secretary problem has demonstrated that the optimality of candidate-counting algorithms. These are defined as algorithms that set exploration periods – defined by the number of candidates seen – after which any candidate that appears that is better than all candidates already seen may be selected.

We further refine this definition by specifying that we are working with static algorithms, whose exploration periods have defined sizes and do not change as the algorithm runs. Finally, we define simple algorithms as a class of algorithms which sets a single exploration period for each type of candidate after which the algorithm begins hiring the candidates who are better than all other candidates seen so far, or contenders. This class of algorithms is attractive for its simplicity and near-optimal behavior.

---

1In addition, experimental evidence in the paper “In Sequential decision making with relative ranks: An experimental investigation of the secretary problem” suggests that when human evaluators are presented with such an utility function on the Single Secretary problem, their decisions are best explained with this type of algorithm.
When we examine affirmative action mechanisms, we preserve a simple group-blind selection mechanism.

**Group Labels.**

Each of the candidates belongs to either the *majority* group or the *minority group*. The pool of candidates has a fraction $p$ minority candidates and $1 - p$ majority candidates. We make the simple assumption that $p < .5$. Labels are chosen uniformly randomly from the $p^n$ possible different combinations. Thus, the true ordinal ranking of the candidates is uncorrelated with the group labels. The probability that the best candidate is from the minority group is therefore $p$, and the probability that the best candidate is from the majority group is $1 - p$.

**The Form of the Bias.**

Bias changes how the true ranking of the candidates is perceived by the observer. We will specify two assumptions for our form of bias.

1. It is unidirectional and negative against minority candidates. That is to say, if bias exists, it causes minority candidates to be perceived as lower-ranked than they would be in the true rankings.

2. Bias preserves the order among the minority candidates.

We characterize our level of bias as a ceiling $j$, where $j$ specifies the highest ranking of any minority candidate in the perceived ordering. We can interpret a $j = 5$ as “no minority candidate will appear higher than 5th in the perceived ranking” and equivalently as “there will be at least four candidates higher than the highest ranked minority candidate.” Furthermore, we specify that if the true top candidate is minority and there is bias, the observer will perceive exactly four candidates ranked higher than the top minority candidate.

For the purposes of analyzing the expected probability of success of a strategy and the error rates on when the best candidate is minority or majority, we argue that for the any form of non-random form of bias, the effect can be closely approached by a $j$.

Assume a non-random form of bias that is unidirectional and negative against minority candidates that also preserves the relative rank of minority candidates. Let $j$ be the rank that the best minority candidate would appear to the observer if the best minority candidate were the best overall candidate. For example, if we assumed a constant form of bias, such as “add 4 to the rank of every candidate and reorder” this would be characterized as $j = 5$. If we assumed a multiplicative form of bias – “multiply the ranks of the every minority candidate by 10” – we could still characterize this as $j = 10$ for our analysis. We argue that the bias applied to every other candidate besides for the overall top candidate when they minority is largely extraneous to our analysis.

To prove this, we begin with our simple strategies which don’t include affirmative action policies and are group-blind. Because they are group-blind, their success rate depends only on the arrival order of the perceived rankings, which is unchanged. Consider that uniformly random distribution of rankings and labels. Assume that the best candidate is a majority candidate. Now, for any given location in the arrival order, whether bias exists or not, all perceived rankings of the candidates that arrive will be equally likely. This is a consequence of the uniform random distribution of the labels. Bias might affect the distribution of group labels in the candidates selected, but it will not affect our probability of selecting the best candidate. Furthermore, because it does not affect this probability, it will not affect the error rates when the best candidate is minority or majority.

Now, we make use of the second specification: the preservation of relative order among minority candidates. We examine two possible affirmative action policies: a *window* policy and a *designated slot* policy. The window policy depends only on the relative ranking of minority candidates, so our analysis still applies. In the case of the designated-slot policy, this policy *does* depend on the amount of bias
applied to other candidates. However, we can bound the probability that this has an effect by the term $pq \cdot (1 - C)^{j+1}$, which tends to be very small.\(^2\)

**The Existence of the Bias.**

The probability that bias exists is denoted by probability $q$.

### 2.2 Notation

We let $k$ denote the number of selections that we can make, $p$ denote the proportion of minority candidates, $j$ denote the minimum rank of the highest minority candidate if there is bias, and $q$ be the probability that there is bias.

We will use the term “contenders” to refer to candidates that appear to be more highly ranked than all other candidates that have appears before them. This refers only, however, to the perceived ranking of the candidates rather than the true ranking. The phrase “minority contender” denotes a minority candidate who appears more highly ranked than all minority candidates already seen, rather than “a contender who is from the minority group.”

We refer to the assumption that $q = 0$ as the utopian assumption and to selection mechanisms which do not examine the group of candidates as *group-blind*.

We denote the expected success rate of the group-blind algorithm and the two affirmative action mechanisms with:

- Group-Blind Strategy: $W(K, T)$
- Designated Slot Mechanism: $W(K - 1, 1, T, C)$
- Window Mechanism: $W(K, T, T_C)$

where $K$ denotes the number of candidates\(^3\), $T$ always denotes the exploration period of the group-blind selection mechanism, and $T$ and $T_C$ will refer to the exploration periods of our affirmative action mechanisms.

We will use $X(T)$ to denote the probability that a contender (or a minority contender, depending on the strategy) that is accepted by a selection mechanism is the best overall candidate.

Finally, we will continue to examine the same single example throughout the paper, using $k = 7$ and $p = 0.20$ as our standard scenario. In addition, we choose to examine three levels of bias throughout, denoting $j = 3$, $j = 6$, and $j = 10$ as “low”, “medium”, and “high” levels of bias.

### 3 Optimal Group-Blind Solutions

#### 3.1 Analysis of the Optimal Algorithm

We begin by presenting the optimal static and simple algorithm for the multi-choice iteration of the secretary problem as described in earlier literature.\(^{[15]}\) The true optimal strategy includes $k$ starting numbers corresponding to $k$ different sizes for exploration periods after which the next contender after that exploration period is accepted. Gilbert and Mosteller, however, demonstrated that the success rate of the simple algorithm very closely approximates the success rate of the optimal algorithm.

\(^2\)This term will become more clear in Chapter 4, when we have examined the strategies, but essentially, if the group-blind algorithm selects a minority contender that the designated would have selected that is not the best minority candidate, this level of bias will have had an effect. However, this term is presented in Appendix B an upper bound for the probability that this will occur. In our analysis, $C$ is lower bounded at $\frac{1}{e}$, so this entire term converges to 0 extremely quickly.

\(^3\)We use capital $K$ when denoting the parameter in the expected success function and lowercase $k$ in the formulas.
The analysis of the optimal simple strategy proceeds with two steps.

(A) We calculate the asymptotic probability distribution of the number of contenders\(^4\) in a connected set of interviews.

(B) We apply this result to calculate the asymptotic probability of winning the \(k\)-choice game with the strategy that passes the first \(T\) draws and chooses the next \(k\) contenders, if they exist.

We begin with Step A and the asymptotic probability distribution for the number of contenders.

**Lemma 3.1.** The probability of exactly \(k\) candidates appearing from candidates between a set of interviews for which \(a < i \leq b\) is given asymptotically \((n \to \infty)\) by \(p(k|a, b)\) where:

\[
p(k|a, b) = \left(\frac{a}{b}\right) \left(\frac{\log\left(\frac{b}{a}\right)}{k!}\right)^k
\]

We provide the first three cases and prove the general case through induction. Throughout our analysis, we assume that \(a\) and \(b - a\) are far greater than \(k\).

- **\(k = 0\).** Notice that if the best candidate of the first \(b\) candidates appears before the \((a + 1)\)th interview, we will have 0 contenders in our set of interviews, so \(p(0|a, b) = \frac{a}{b}\).
- **\(k = 1\).** If there is only one candidate, the candidate must be the best of the first \(b\) candidates (occurs with probability \(\frac{1}{b}\)), and the second best of the first \(j - 1\) draws must be in the first \(a\) draws (occurs with probability \(\frac{a}{j - 1}\)). Summing over the possible values of \(j\), we have:

\[
p(1|a, b) = \sum_{j = a + 1}^{b} \frac{a}{b(j - 1)}
\]

\[
\approx \left(\frac{a}{b}\right) \log\left(\frac{b - 1}{a}\right)
\]

\[
\approx \left(\frac{a}{b}\right) \log\left(\frac{b}{a}\right)
\]

- **\(k = 2\).** The \(j\)th draw \((a < j \leq b)\) is the second and last candidate in a set if it is the largest in the first \(b\) candidates, and there is exactly 1 candidate between \(a\) and \(j - 1\): namely, the above equation, if we replace \(b\) with \(j - 1\). We sum this product over the possible values of \(j\) and get:

\[
p(2|a, b) = \sum_{j = a + 2}^{b} \frac{a \cdot \log\left(\frac{j - 1}{a}\right)}{b(j - 1)}
\]

\[
\approx \left(\frac{1}{b}\right) \int_{a + 1}^{b} \left(\frac{a}{x}\right) \log\left(\frac{x}{a}\right) dx
\]

\[
\approx \left(\frac{a}{b}\right) \frac{\log\left(\frac{b}{a}\right)^2}{2!}
\]

Now, to complete the induction, we assume that \(p(k|a, b) \approx \left(\frac{a}{b}\right) \left(\frac{\log\left(\frac{b}{a}\right)}{k!}\right)^k\), and we use the same argument: the \(j\)th draw is the last – and the \((k + 1)\)th candidate if it is the largest among the first \(b\) draws, and \(k\) candidates appear in the interval from \(a + 1\) to \(j - 1\). Therefore, we get the probability at a given \(j\) from \(p(k|a, k - 1, \frac{1}{b})\) and then we sum and approximate as we did going from \(k = 1\) to \(k = 2\).

\(^4\)Recall that a “contender” refers to a candidate who is better than all of the candidates that have been seen so far.
Next, we calculate the optimal exploration period. Let the number of candidates that appear be $\omega$. If $\omega = 0$ or $\omega > k$ candidates, we lose. Otherwise, if $\omega \in [1,k]$ then our algorithm will have succeeded. Thus, Lemma 3.1 implies the following:

**Lemma 3.2.** The probability of selecting the best candidate from a pool of $n$ candidates with $k$ selections is given by the following probability

$$\binom{T}{n} \sum_{i=1}^{k} \frac{\left[ \log \left( \frac{n}{T} \right) \right]^i}{i!}$$

To optimize the winning probability, we let $T = \alpha n$ and let $\alpha = e^{-a}$. This gives us the expression:

$$P(\text{win with } k \text{ choices starting after } T) = e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!}$$

We differentiate this with respect to $a$ in order to solve for the maximum success probability, and we get that $a^* = (k!)^{\frac{1}{k}}$. This gives us the probability of winning $e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!}$, and the corresponding values are denoted below.

## The Optimal Simple Strategy

<table>
<thead>
<tr>
<th>Choices</th>
<th>Success Probability</th>
<th>$a^*$</th>
<th>$e^{a^*}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.368</td>
<td>1.000</td>
<td>0.368</td>
</tr>
<tr>
<td>2</td>
<td>0.587</td>
<td>1.414</td>
<td>0.243</td>
</tr>
<tr>
<td>3</td>
<td>0.726</td>
<td>1.817</td>
<td>0.162</td>
</tr>
<tr>
<td>4</td>
<td>0.817</td>
<td>2.213</td>
<td>0.109</td>
</tr>
<tr>
<td>5</td>
<td>0.877</td>
<td>2.605</td>
<td>0.074</td>
</tr>
<tr>
<td>6</td>
<td>0.917</td>
<td>2.994</td>
<td>0.050</td>
</tr>
<tr>
<td>7</td>
<td>0.944</td>
<td>3.380</td>
<td>0.034</td>
</tr>
<tr>
<td>8</td>
<td>0.962</td>
<td>3.764</td>
<td>0.023</td>
</tr>
<tr>
<td>9</td>
<td>0.974</td>
<td>4.147</td>
<td>0.016</td>
</tr>
</tbody>
</table>

**Figure 1.**

### 3.2 An Alternative Approach

Now, we demonstrate an alternative approach to calculating the same optimal value of $a^*$. Rather than directly optimizing the win probability, we show that it is equivalent to calculate a $T$ such that the marginal net utility of accepting a contender that appears as the $T^{th}$ candidate is positive. Essentially, we decompose the marginal net utility into the marginal benefit and the marginal cost of accepting such a contender, and demonstrate the intuition that we maximize utility when these terms are equal.

Recall that $W(K,T)$ denotes the asymptotic probability of selecting the best overall candidate with a simple strategy that begins selection after $T$ candidates have been seen. In addition, recall that $X(T)$ denotes the chance that a contender which appears as the $T^{th}$ candidate is the best overall candidate.

**Theorem 3.3.** $T^*$ maximizes $W(K,T)$ if $T^*$ solves

$$X(T^*) - (W(K,T^*) - W(K-1,T^*)) = 0$$

**Proof.** First, we note that $W(K,T)$ is defined over $K \in \mathbb{N}$ and $T \in [0,1] \cap \mathbb{Q}$. Thus, our partial derivative with respect to $T$ is defined over the rationals, and we express $T$ as $\frac{a}{b}$ and $h$ as $\frac{c}{d}$.
\[
\frac{\delta}{\delta T} W(K, T) = \lim_{h \to 0} \frac{W(K, T - h) - W(K, T)}{h}
\]
\[
= \lim_{z \to 0} \frac{W(K, \frac{ad - cb}{bd}) - W(K, \frac{ad}{bd})}{z}
\]
\[
= \lim_{z \to 0} \frac{d}{c} \cdot \left( W(K, \frac{ad - cb}{bd}) - W(K, \frac{ad}{bd}) \right)
\]

Now, because both \(W(K, \frac{aN - 1}{bN})\) and \(W(K, \frac{aN - 1}{bN})\) are rational numbers expressed with the same denominator, we can directly compare the performance of these two strategies. They differ in only one way – \(W(K, \frac{ad - cb}{bd})\) begins selecting contenders \(cb\) candidates before \(W(K, \frac{ad}{bd})\). Thus, the performance of these strategies is the same unless contenders are selected by the former strategy from these \(cb\) candidates. As in Lemma 3.1, we let \(p(i | a, b)\) be the probability of finding \(i\) contenders between \(a\) and \(b\). We let \(\theta(i, \frac{a_{\text{total}}}{b_{\text{total}}} = \frac{a}{b})\) be the probability of any of those \(i\) contenders selected being the best.

\[
= \lim_{z \to 0} \frac{d}{c} \cdot \left( \sum_{i=0}^{K} p(i | ad - cb, ad) \cdot \left( W(K - i, \frac{a}{b}) + \theta(i, \frac{ad - cb}{bd} + \frac{ad}{bd}) \right) \right) - \frac{d}{c} \cdot W(K, \frac{a}{b})
\]
\[
= \lim_{z \to 0} \frac{d}{c} \left( \sum_{i=0}^{K} \frac{ad - cb}{ad} \cdot \frac{\ln(\frac{ad}{ad-cb})}{i!} \left( W(K - i, \frac{a}{b}) + \theta(i, \frac{a}{b} - \frac{c}{d} \frac{b}{a}) \right) \right) - \frac{d}{c} \cdot W(K, \frac{a}{b})
\]

This probability converges to 0 for all \(i > 0\). The only non-zero probabilities are for \(i = 0\) and \(i = 1\) which are given below:

\[
i = 1 \implies \lim_{z \to 0} \frac{d}{c} \cdot \left( 1 - \frac{cb}{ad} \right) \cdot \ln \left( \frac{1}{1 - \frac{cb}{ad}} \right) = \frac{b}{a}
\]
\[
i = 0 \implies \lim_{z \to 0} \frac{d}{c} \cdot \left( 1 - \frac{cb}{ad} \right) = \left( \frac{d}{c} - \frac{b}{a} \right)
\]

Thus, our derivative can be rewritten as below:

\[
= \lim_{z \to 0} \left( \frac{d}{c} - \frac{b}{a} \right) \cdot W \left( K, \frac{a}{b} \right) + \frac{b}{a} \left( W(K - 1, \frac{a}{b}) + \theta(1, \frac{a}{b} - \frac{c}{d} \frac{a}{b}) \right) - \frac{d}{c} \cdot W \left( K, \frac{a}{b} \right)
\]
\[
= \lim_{z \to 0} \frac{b}{a} \cdot \left( W(K - 1, \frac{a}{b}) - W(K, \frac{a}{b}) + \theta(1, \frac{a}{b} - \frac{c}{d} \frac{a}{b}) \right)
\]
\[
= \lim_{z \to 0} \frac{b}{a} \cdot \left( W(K - 1, \frac{a}{b}) - W(K, \frac{a}{b}) + \theta(1, \frac{a}{b}, \frac{a}{b}) \right)
\]

We note that \(\theta(1, \frac{a}{b} - \frac{c}{d} \frac{a}{b})\) was the probability that the contender we select after seeing between \(\frac{a}{b} - \frac{c}{d} \frac{a}{b}\) of the candidates and \(\frac{a}{b}\) of the candidates is the best overall candidate. This probability is continuous in the fraction of candidates seen and therefore:

\[
\lim_{z \to 0} \theta \left( 1, \frac{a}{b} - \frac{c}{d} \frac{a}{b} \right) = \theta \left( 1, \frac{a}{b}, \frac{a}{b} \right) = X \left( \frac{a}{b} \right)
\]

Thus, rewriting \(z\) as \(T\), we arrive at:

\[
\frac{\delta}{\delta T} W(K, T) = \frac{1}{T} \left( W(K - 1, T) - W(K, T) + X(T) \right)
\]
Finally, the $T^*$ which maximizes $W(K,T)$ must satisfy
\[
\frac{1}{T} (W(K-1,T) - W(K,T) + X(T)) = 0
\]
\[
X(T) - (W(K,T) - W(K-1,T)) = 0
\]
Thus, we have proven our theorem.
We check now to make sure that the results of this theorem yield the same optimal values as calculated in Figure 1. First, we calculate $X(T)$.

**Lemma 3.4.** In this unbiased scenario, $X(T) = \frac{T}{N}$

**Proof.** In the unbiased case, we simply note that each candidate has a $\frac{1}{N}$ probability of being the best overall candidate. Thus, given a group of $\frac{T}{N}$ candidates chosen uniformly randomly from all possible groups, there is a $\frac{T}{N}$ probability that the best overall candidate is within this group. In addition, for any group of ranks, the probability that the best candidate of that group is the last in the arrival order is independent of the ranks that the group contains. Thus, conditioned on the fact that candidate $T$ is a contender and we are selecting the best candidate of this group, $X(T) = \frac{T}{N}$.

Next, we have that $W(K,T) - W(K-1,T)$ is simply the probability of winning with the $K^{th}$ selection, which is described in Lemma 3.1 as $e^{-\alpha} \left( \frac{a^K}{K!} \right)$

Thus, solving $X(T) = W(K,T) - W(K-1,T)$ and writing $\frac{T}{N} = e^{-\alpha}$, we solve $e^{-\alpha} = e^{-\alpha} \left( \frac{a^K}{K!} \right)$ to arrive at $a = k!^{\frac{1}{2}}$ as we desired.

4 Cost of Implicit Bias

The optimal algorithm as described in Section 3 assumed that there was no bias and did not take group labels into account. We refer to the first assumption as the utopian assumption and to algorithms which do not look at group labels as “group-blind”. Now, we drop the utopian assumption and we examine the performance of group-blind algorithms in biased environments.

4.1 An Upper Bound for Group-Blind Algorithms

**Theorem 4.1.** Any “group-blind” algorithm, given an implicit bias $j$ has the following upper bound on its performance:

\[
W(K,T) \leq (1 - p) \cdot \frac{T}{N} \sum_{i=1}^{K} \left( \ln\left( \frac{N}{T} \right) \right) + p \cdot \frac{1}{j} \prod_{i=0}^{j-1} \left( \frac{N - T - i}{N} \right)
\]

Let us denote the chance of finding the best candidate with a group-blind strategy as $P(1^{st}|T,K)$. In addition, our form of bias specifies that the best candidate when from the minority group will appear as the $j^{th}$ highest rank in the observed rankings of the evaluating agent. Thus, we will denote the probability of finding the $j^{th}$ best applicant from the observed rankings with a starting point of $T$ as $P(j^{th} | K, T)$.

Because the group labels are drawn uniformly randomly from all possible combinations, the best candidate is from the minority group with probability $p$ and from the majority group with probability $1 - p$. This means that the expected performance of our algorithm with stop time $T$ is bounded above by:

\[
(1 - p) \cdot P(1^{st}|T,K) + p \cdot P(j^{th}|T,K)
\]
The first term has already been computed from Lemma 3.2, and so now we construct an upper bound on $P(j^{th}|T,K)$ by noting that our algorithm only selects contenders. For the best minority candidate to be a contender with $j$ candidates who appear higher ranker, our best minority candidate must appear before all of them. In addition, in order to be “selectable”, $j$ must appear after the exploration period of $T$ candidates. Now, the probability that all of them fall after the first $T$ candidates is

$$\prod_{i=0}^{j-1} \left( \frac{N - T - i}{N} \right)$$

and we multiply all of this by $\frac{1}{j}$, which is the probability that of these top $j$ candidates, the $j^{th}$ highest appears before the rest. This probability does not factor in the probability that more than $K$ contenders may appear before the $j^{th}$ highest candidate but rather assumes that if the $j^{th}$ highest appears as a contender, they will be selected by our algorithm. Thus, this probability serves as an upper bound.

$$P(j^{th} | T, K) \leq \frac{1}{j} \prod_{i=0}^{j-1} \left( \frac{N - T - i}{N} \right)$$

Thus, we upper bound the performance of a simple strategy with:

$$\text{Expected Utility} \leq (1 - p) \cdot \frac{T}{N} \sum_{i=1}^{K} \left( \frac{\ln(N)^i}{i!} \right) + p \cdot \frac{1}{j} \prod_{i=0}^{j-1} \left( \frac{N - T - i}{N} \right)$$

We note that for a small and finite $j$ as $N \to \infty$,

$$\lim_{N \to \infty} \prod_{i=0}^{j-1} \frac{N - T - i}{N} \to \left( 1 - \frac{T}{N} \right)^j$$

Therefore, rewriting $\frac{T}{N}$ as $e^{-a}$, we can upper bound our expected utility with:

$$\text{Expected Utility} \leq (1 - p) e^{-a} \sum_{i=1}^{K} \frac{\ln(N)^i}{i!} + p \cdot \frac{1}{j} \left( 1 - \frac{T}{N} \right)^j$$

$$\leq (1 - p)e^{-a} \sum_{i=1}^{K} \frac{a^i}{i!} + p \left( 1 - e^{-a} \right)^j$$

We solve for the form of the derivative and set it equal to 0.

$$\frac{\delta}{\delta a} (\text{Expected Utility}) = (1 - p) \left( e^{-a}(1 - \frac{a^k}{k!}) \right) + pe^{-a}(1 - e^{-a})^{j-1}$$

$$0 = (1 - p) \left( 1 - \frac{a^k}{k!} \right) + p \cdot (1 - e^{-a})^{j-1}$$

This implicit equation can be solved for specific values of $p$, $k$, and $j$. As introduced in Section 2, in the following analysis we will examine our standard scenario where $k = 7$ and $p = 0.20$.

First, however, we will show that the assumption that we made in solving for our upper bound – in the case that if the best candidate were minority, if they were in a position that they would appear as a contender than they would be selected – does not significantly impact our value. In short, we show that this upper bound is fairly tight.
Lemma 4.2. The optimal group blind algorithm differs from the optimal upper bound calculated in Theorem 3.1 by at most a factor of:

$$\frac{p}{j} \left[ 1 - \left( 1 - e^{-K(\frac{1}{j})^{j-1}} \right)^{j-1} \right] \cdot e^{-K(\frac{1}{j})} \sum_{i=1}^{K} \frac{(K!j^i)}{i!}$$

We noted that $\frac{1}{j}(1 - e^{-a})^j$ was an overestimate of the probability that the best overall candidate would be selected in the biased scenario because we assumed that all contenders would be selected by the algorithm. Now, we lower bound this probability. First, recall that $\frac{1}{j}(1 - e^{-a})^j$ denotes the probability that in the biased scenario, the best candidate when minority will appear ahead of all of the higher ranked candidates and beyond the exploration period.

Instead of assuming that we select all contenders, we can multiply this term by the probability that the highest contender will be selected. This is simply $e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!}$, and to see that this is a lower bound, we can simply note that because we know that our best minority candidate will appear as a contender, this is the same problem as the unbiased problem on a smaller subset of the candidate.

First, we use the fact that we are sure that the best minority candidate will appear after the exploration period to condition on that probability – $e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!}$ multiplied together, this gives us $\frac{1}{j}(1 - e^{-a})^{j-1} \cdot e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!}$.

Furthermore, we can confirm this is a lower bound by noting that we could condition the location of this best candidate on the fact that it is the first of the $j$ highest candidates to appear, and therefore, rather than a uniform distribution over the slots after the exploration period, we are likely running the multi-choice secretary problem over a smaller group. Therefore, it is more likely to be selected than our bound suggests.

Our lower bound is given by:

$$\max_a (1 - p)e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!} + \frac{p}{j}(1 - e^{-a})^{j-1} \cdot e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!}$$

Now, we show choose $a^* = k!\frac{1}{j}$, the value that optimizes the first term. This must be greater than or equal to the first term in our upper bound.

In fact, we can bound our upper bound again with:

$$\max_a (1 - p)e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!} + \frac{p}{j}(1 - e^{-a})^j \leq \max_a (1 - p)e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!} + \max_a \frac{p}{j}(1 - e^{-a})^j$$

$$\leq \max_a (1 - p)e^{-a} \sum_{i=1}^{k} \frac{a^i}{i!} + \frac{p}{j}$$

Thus, the difference between our upper bound and the lower bound is bounded by

$$\frac{p}{j} \left[ 1 - \left( 1 - e^{-K(\frac{1}{j})^{j-1}} \right)^{j-1} \right] \cdot e^{-K(\frac{1}{j})} \sum_{i=1}^{k} \frac{(K!j^i)}{i!}$$

This is decreasing in $j$ and $k$. For even relatively small values of $k$ and $j$ – namely, where $k \geq 5$ and $j \geq 3$ – we have that this is bounded by .00159 · $p$. When we consider our assumption that $p < .5$, this is always less than a tenth of a percent. Thus, our upper bound is fairly tight.
4.2 Visualizing the Effects of Bias

We work now to visualize the effect of bias on the performance of group-blind algorithms and how this effect responds to the different parameters in our model: \( q \), \( p \), and \( j \).

**Upper Bounds on Success Probabilities**

*with \( k = 7 \)*

<table>
<thead>
<tr>
<th>( p )</th>
<th>No Bias</th>
<th>( j = 3 )</th>
<th>( j = 6 )</th>
<th>( j = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.9436</td>
<td>0.9114</td>
<td>0.9032</td>
<td>0.8999</td>
</tr>
<tr>
<td>0.10</td>
<td>0.9436</td>
<td>0.8793</td>
<td>0.8628</td>
<td>0.8563</td>
</tr>
<tr>
<td>0.15</td>
<td>0.9436</td>
<td>0.8473</td>
<td>0.8225</td>
<td>0.8128</td>
</tr>
<tr>
<td>0.20</td>
<td>0.9436</td>
<td>0.8153</td>
<td>0.7822</td>
<td>0.7692</td>
</tr>
<tr>
<td>0.25</td>
<td>0.9436</td>
<td>0.7833</td>
<td>0.7420</td>
<td>0.7257</td>
</tr>
<tr>
<td>0.30</td>
<td>0.9436</td>
<td>0.7514</td>
<td>0.7018</td>
<td>0.6822</td>
</tr>
<tr>
<td>0.35</td>
<td>0.9436</td>
<td>0.7196</td>
<td>0.6616</td>
<td>0.6388</td>
</tr>
<tr>
<td>0.40</td>
<td>0.9436</td>
<td>0.6878</td>
<td>0.6216</td>
<td>0.5955</td>
</tr>
<tr>
<td>0.45</td>
<td>0.9436</td>
<td>0.6562</td>
<td>0.5816</td>
<td>0.5522</td>
</tr>
</tbody>
</table>

*Figure 2.*

**The Cost of Bias on the Utopian Group-Blind Strategy**

*for \( k = 7, p = .20 \)*

*The error bars represent the 99% confidence intervals*

*Figure 3.*

We first note that Figure 2 works with the upper bounds of success probabilities, while Figure 3 examines the performance of the group-blind strategy when we have optimized under the assumption of no-bias – we will call this the utopian group-blind strategy.
Figure 3 is a striking example of the significant effect of the low levels of bias. Even at $j = 2$, the lowest level of bias which will affect the performance of an algorithm, we can see a significant drop in the performance of the algorithm.

We also observe that the severity of this effect tapers off. Of course, the effect of bias is upper bounded by $p$: this follows from Theorem 3.1 and noting that

$$f_2(x) \in [0, 1] \forall x \implies \max_x f_1(x) \leq \max_x (f_1(x) + f_2(x))$$

This trend, however, might be better understood if we examine the cost in performance as a proportion of $p$, so we show Figure 2 with costs as percentages.

![Table of costs as a percentage of p](image)

We conclude that varying $p$ does not have a significant effect on cost as a percentage of $p$: it is primarily the level of bias for a given $k$ that dictates this proportional loss. Again, the cost of implicit bias is severe: for a “low” level of bias, the optimal group-blind algorithm must perform at least about $\frac{2}{3} \cdot p$.

5 Diversity Policies

We can see from Figure 3 that the effects of very small amounts of bias are extremely large. In addition, Figure 4 has illustrated how this steep loss, as a proportion of $p$, is determined primarily by $j$. In some ways, this is unsurprising – work on a variant of the Secretary Problem has demonstrated the difficult of selecting the second-best candidate even with a strategy that is optimized for doing so (Postdoc Variant).

Given the potentially costly effects of implicit bias, we introduce two types of affirmative action mechanisms. We will then extend Theorem 3.3 to these strategies in order to give us a structure with which to establish theoretical and experimental bounds on their performance, and we will use these to characterize the scenarios in which these mechanisms increase expected utility.

---

5The motivating story presented by Robert Vanderbei was as follows: we are trying to hire a postdoc and we are confident that the best applicant will receive and accept an offer from Harvard. Given $k = 1$, the optimal algorithm is to select the second-best of all candidates who appear after an exploration period. This, however, has an asymptotic success rate of $\frac{1}{e}$ as compared to the $\frac{1}{2}$ that we established on selecting the best overall candidate.

6These mechanisms will involve a combination of general and minority candidate search processes. We examine strategies which fully separate candidates in Appendix C but have excluded them from the main body of this paper because of their lower expected utility.
5.1 The Designated Slot Mechanism

We first propose and analyze a simple correction mechanism: a designated slot which selects a minority contender. This policy is reminiscent of the NFL’s Rooney Rule in a dynamic selection context. We implement this policy as a static and simple strategy and provide the following characterization.

We characterize the parts of this strategy: the $T$ selection mechanism and then the designated slot. The $T$ selection mechanism is simply the group-blind simple strategy that begins selecting contenders after $T$ candidates have been seen and uses up to $K-1$ selections. The designated slot mechanism, on the other hand, selects the first minority contender that would not have been selected by the $T$ selection mechanism after we have seen a fraction $C$ of the minority candidates. We denote the expected success of this policy with $W(K-1,1,T,C)$. Notice that we could allocate more than one designated slot, but we restrict our analysis to the case in which $z=1$.

We begin by providing bounds on the optimal value $T^*$. We use this $T^*$ and a lower bound on the performance of our this strategy to choose a $C^*$, and then we characterize the ranges of $q$ such that our designated slot policy will increase expected utility.

First, we extend Theorem 3.3 to this strategy. The proof is demonstrated in Appendix E, but we essentially demonstrate that we may divide out the probability that the $T$ and $C$ selection mechanisms would choose the same candidate in isolation.

**Theorem 5.1.** For any given $C$, the $T^*$ that maximizes $W(K-1,1,T,C)$ must also solve:

$$X(T^*) - (W(K-1,1,T^*,C) - W(K-2,1,T^*,C)) = 0$$

In order for this to be useful, we must be able to characterize both terms in this equation: $X(T)$ and $W(K-1,1,T^*,C) - W(K-2,1,T^*,C)$. We bound their possible values below.

**Bounding $X(T)$.**

Recall that $X(T)$ denotes the probability that a contender selected as the $T^{th}$ candidate seen is the best overall candidate. In the cases in which there is no bias or the top candidate is from the majority group, the probability that a contender selected as the $T^{th}$ candidate is the best overall is given by the fraction $T/N$.

The expected utility in the unbiased case is simply given by Lemma 3.4: $X(T) = T/N$.

In the case where bias exists and the best overall candidate is a majority candidate, we simply apply the argument presented in Section 2. The bias does not change the perceived ranking of the best overall candidate, and our selection at $T^*$ is group-blind. Therefore, we argue that all possible arrival orders of perceived rankings are still equally likely. Thus we may argue that the group of the first $T/N$ candidates is still chosen uniformly randomly from all possible groups, and therefore the best candidate of this group has a $T/N$ probability of being the best overall.

For the last case, however, in which the best candidate is from the minority group and bias exists, we may lower bound and upper bound the probability that our candidate is the best overall. Given that $X(T)$ is the probability that a contender who appears as the $T^{th}$ candidate to be seen is the best overall, we first compute the probability that the best overall candidate appears as a contender due to bias. We use the same probability that we calculated in Section 3: $\frac{1}{j}(1 - \frac{T}{N})^j$. Next, we apply the same logic as we did in Section 3 to create a lower and upper bound.

Conditioned on appearing at the best candidate appearing in the $T^{th}$ slot or after but also before the $j$ higher candidates in the perceived rankings, the probability distribution of the location of the best

---

7 Recall that in Section 2, we defined “minority contender” as a minority candidate that is more highly ranked than all other minority candidates.

8 A policy implemented in the American NFL in 2003 that requires that at least one of the finalists for every head coaching and senior operations opening be chosen from a minority group.
overall candidate is no longer uniform over \([T,N]\) but rather skewed towards earlier arrival. This makes it more likely that the best overall candidate appears in the \(T^{th}\) candidate slot than in the uniform distribution. Thus, we can lower bound the probability that the \(T^{th}\) candidate is the best overall with the probability suggested by the uniform distribution, which gives us \(X(T) = \frac{1}{2}(1 - \frac{T}{N})^j \cdot \frac{1}{N^j}\). We can use the probability that the best overall candidate simply appears as a contender as an upper bound. \(X(T) = \frac{1}{2}(1 - \frac{T}{N})^j\).

Let us rewrite \(\frac{T}{N}\) as \(e^{-a}\). We have that

\[
e^{-a} \cdot \left[ (1 - pq) + pq \cdot \frac{1}{j} (1 - e^{-a})^j \right] \leq X(e^{-a}) \leq e^{-a} (1 - pq) + pq \cdot \frac{1}{j} (1 - e^{-a})^j
\]

Bounding \(W(K - 1, 1, T^*, C) - W(K - 2, z, T^*, C)\).

We may describe this term as the probability that having an additional \(K - 1^{st}\) slot is the difference in success and not. We examine three scenarios and bound them separately.

The \(1 - p\) scenario. In the scenario that the best candidate is from the majority group, we may simply characterize this difference as the probability that we select the best candidate with the \(K - 1^{st}\) selection. This is the probability of there being exactly \(K - 1\) candidates, which is given by Lemma 3.1. Thus, this contributes:

\[
(1 - p) \cdot e^{-a} \frac{a^{K-1}}{K - 1!}
\]

The \(p \cdot (1 - q)\) scenario. In the scenario that the best candidate is from the minority group but there is no bias, we may lower bound this difference as the probability that we select the best candidate with the \(K - 1^{st}\) selection. However, there is the probability that this \(K - 1^{st}\) candidate selects a minority candidate, which enables the designated slot mechanism to pass on a minority contender and select the correct candidate. We denote this event that our algorithm selects a minority candidate with its \(K - 1^{st}\) selection as \(P_{K-1}\) and the probability that this difference lets the designated slot mechanism succeed (conditioned on \(P_{K-1}\)) is \(DS_{K-1}\). Thus, we can add the quantity

\[
p \cdot (1 - q) \cdot e^{-a} \frac{a^{K-1}}{K - 1!} + p \cdot (1 - q) + y
\]

such that \(y \in [0, P(1_P_{K-1}) \cdot DS_{K-1}]\)

The \(p \cdot q\) scenario. We divide this scenario into two parts as well. First, we examine the probability that the best candidate, although bias is working against them, is directly selected by the algorithm as its \(K - 1^{st}\) selection. We lower bound this scenario with the probability \(0\) and upper bound this with the familiar term \(\frac{1}{j} (1 - e^{-a})^j\). Next, we examine the scenario that in the biased scenario, our algorithm again selects a minority candidate with its \(K - 1^{st}\) selection as \(P_{K-1}\). We lower bound this with \(0\) and we can upper bound it with \((1 - C)^j\).\(^9\)

Thus, in this scenario, we have that the probability that the \(K - 1^{st}\) candidate makes the difference between succeeding and not succeeding is \(pq \cdot (z_1 + z_2)\) where:

\(^9\)We note that we may upper bound the probability that any of our \(K - 1\) selections on is used on a minority contender after seeing the fraction \(C\) of the minority candidates. For a minority contender better than all candidates to appear, we must not have seen any of the \(j - 1\) majority candidates who appear higher ranked than all of the minority candidates. The probability that we don’t see any of them until we have pulled a fraction \(C\) of the minority candidates is given by \((1 - C)^{j-1}\) which is proved in Appendix B. This probability is independent from the probability that there are more than 0 minority contenders – defined only in minority candidates, which is \(1 - C\). Thus, the product of these two probabilities gives us an upper bound of \((1 - C)^j\)
Now, having established these bounds, we choose sample values with which to compute a theoretical lower bound. We choose the values:

\[ X(e^{-a}) = (1 - pq + pq \cdot \frac{1}{j}(1 - e^{-a})^j) \cdot e^{-a} \]

\[ W(K - 1, 1, T, C) - W(K - 2, 1, T, C) = (1 - pq) \cdot P(K - z, T) + pq \cdot \frac{1}{j}(1 - e^{-a})^j \]

Thus, we arrive an estimate of \( T^* = e^{-(k - z)!/(k-z)} \).

**Lower Bounding the Expected Utility.**

Now, we examine how we are going to lower bound \( W(K - 1, 1, T^*, C) \). We provide this lower bound below before breaking down each of the scenarios with similar principles to those established above:

Let \( \frac{C}{N} = e^{-c} \). We claim the following:

**Theorem 5.2.** The designated slot algorithm with \( K \) selections and exploration periods of size \( T \) and \( C \) for the group-blind and designated slot selection mechanisms has an expected utility that is lower-bounded as follows:

\[
W(K - 1, 1, T, C) \leq (1 - p) \cdot e^{-a} \sum_{i=1}^{k-1} \frac{a^i}{i!} + p \cdot (1 - q) \cdot \left[ e^{-a} \sum_{i=1}^{k-1} \frac{a^i}{i!} + f(T, C) \right] + p \cdot q \cdot \left[ e^{-a} \sum_{i=1}^{k-1} \frac{a^i}{i!} \cdot \frac{1}{j}(1 - e^{-a})^j + e^{-c} \cdot c \cdot (1 - (1 - e^{-c})^{j-1}) \right]
\]

*The 1 - p case.* This case follows the logic above, and the designated slot window does not improve the expected utility because it will not select the best candidate, nor will it free up any more selections by the group-blind selection mechanism.

*The p · (1 - q) case.* We lower bound the probability of success by examining the case in which the group-blind selection mechanism succeeds, and we add \( f(T, C) \), which we use to denote the probability that the group-blind selection mechanism fails to find the best candidate but the designated slot mechanism succeeds.

The probability that the designated slot mechanism succeeds conditioned on the failure of the group-blind selection mechanism is difficult to directly compute, but we lower bound this probability by selecting an example for which the probabilities are independent. We consider the scenario in which all \( k - 1 \) selections from the group-blind selection mechanism are used up before the exploration period for the designated slot mechanism is over. We denote the indicator variable for this event with \( \mathbf{1}_{\geq K-1,C} \)

Now, we claim that this probability is independent from the probability that the best candidate is selected by the designated slot mechanism because the number of candidates in one interval of contender is independent with the number of contenders in a non-overlapping interval.
We outline a simple proof. Consider any intervals \([a, b]\) and \([c, d]\) such that \(c > b\). We note that the number of contenders\(^{10}\) in interval \([a, b]\) is not dependent on the absolute ranks of the candidates within \([a, b]\) but rather the relative ordering of the candidates. Therefore, all sets of possible rankings for the candidates in interval \([a, b]\) are equally likely. Therefore, we cannot have gained any information on the set \([c, d]\) from observing the number of contenders in \([a, b]\). Likewise, the observation of contenders in \([c, d]\) can tell us information about the absolute ranks of \([1, c - 1]\), but not information about the relative ranks of the candidates, and the number of contenders in \([a, b]\) only depends on the relative ranks.

Now, \(\mathbb{1}_{k-1} \cdot C\) depends on how many candidates that we have drawn until we have seen \(C\) candidates. Rather than sum over the distribution, we exploit convexity and use a looser lower bound, which turns out to be the probability that there are \(\geq k + 1\) candidates (and for \(k - 1 = 4\), we use the probability that there are over 7 candidates). We relegate this analysis to Appendix Section A.

Finally, we multiply our lower bound for \(\mathbb{1}_{k-1} \cdot C\) by the probability of success for the designated slot mechanism.

Thus, we formally define \(f(T, C)\) for \(k \in [6, 10]\) as follows:

\[
e^{C-T} \cdot \sum_{i=k+1}^{\infty} \frac{(T-C)^i}{i!} \cdot e^{-C}.
\]

and for \(k = 5\), we simply sum from \(i = k + 2\).

The \(p \cdot q\) case. Finally, we use the usual lower bound that we established in Section 3 for the probability that the group-blind selection mechanism chooses the best candidate. To this, we add the probability that the group-blind mechanism fails and the designated slot mechanism succeeds. We characterize this scenario as follows: one of the \(j - 1\) majority candidates which appears more highly ranked than every minority candidate appears while we are still in the designated slot exploration period. The probability that this does not occur is given by \((1 - e^{-c})^{j-1}\) as we establish in Appendix Section B. Thus, the probability that it does occur is given by \(1 - (1 - e^{-c})^{j-1}\). The location of these majority candidates is naturally independent from the location of the best candidate, and we multiply this by the probability that the best candidate appears after the exploration period. This is given by \(1 - C\), and the fact that best candidate appears after one of the \(j - 1\) majority candidates means that the group-blind mechanism will never select it.

We multiply this by the probability that our designated slot mechanism succeeds. Conditioning on the fact that one of the \(j - 1\) majority candidates appears during the exploration period means simply that the group-blind algorithm will not select any minority candidates after the exploration period — otherwise, the success of designated slot mechanism depends only on the relative order of the minority candidates. Conditioning on the fact that the best minority candidate appears after the exploration period, we conclude that there will be \(\geq 1\) contenders. We arrive at:

\[
= (1 - (1 - e^{-c})^{j-1}) \cdot (1 - e^{-c}) \cdot \frac{e^{-c} \cdot C}{1 - e^{-c}}
= (1 - (1 - e^{-c})^{j-1}) \cdot e^{-c} \cdot C
\]

Choosing our \(C^*\).

Now, given our \(T^*\), we would like to choose a \(C^*\) to optimize this lower bound. The term which contains \(C^*\) in the \(p \cdot (1 - q)\) scenario is often negligible\(^{11}\), so we optimize \(C^*\) in the \(p \cdot q\) scenario.

---

\(^{10}\) We ignore the convention of an exploration period here and define a contender as a candidate that is ranked more highly than all candidates previously observed.

\(^{11}\) Optimizing with \(f(T, C)\) with respect to \(C\) separately, we find that \(f(T, C)\) peaks at two-tenths of a percent.
Using this characterization, we calculate a theoretical upper bound for $q$ that is necessary for the lower bound on our designated slot strategy to outperform the optimal group-blind strategy and present our results below.

Figure 5.

Figure 6.

Figure 7.
5.2 Window Policy

Rather than designating a particular slot to minority candidates, we might want to allow the processes to share candidates. In some sense, this means we won’t have a “wasted” slot – if the group-blind algorithm needed to select a candidate and the designated-slot mechanism had already seen the best minority candidate in the exploration period, we might imagine that it would improve expected utility if the group-blind algorithm could “grab” this slot. On the other hand, this might hurt expected utility if the group-blind algorithm grabs this slot when the designated-slot mechanism would’ve otherwise succeeded.

We combine this intuition with another: given that we would like to search over all candidates in a single process, we notice that there are only four types of candidates that we might consider selecting.

- A: Majority candidates that are better than all candidates already seen
- B: Minority candidates that are better than all candidates already seen
- C: Minority candidates better than all minority candidates already seen, but not better than all candidates overall
- D: Majority candidates better than all majority candidates already seen, but not better than all candidates overall

Now, notice that we would never want to take candidates from Category D: these candidates cannot be the best in either the biased or unbiased scenario. The other three categories, however, can be viewed as three simultaneous search processes that we would like to set “thresholds” for.

We take contenders from Category A and B in our simple group-blind strategy, but now we consider candidates of type C. To this end, we propose and analyze another selection mechanism. After we see $T_C$ of the minority candidates, we begin accepting minority candidates of Category C. We denote this strategy which shares the $K$ slots as $W(K,T,T_C)$. In order to analyze this strategy, we bound the possible relationships between the $T$ and $T_C$ exploration periods and choose a relationship from within this bounds. Given this relationship, we will then apply Bayesian Optimization to estimate the optimal $T$ and $T^*_C$.

We again extend Theorem 3.3 to characterize the relationship between the size of these exploration periods.

**Theorem 5.3.** $T^*$ and $T^*_C$ maximize $W(K,T,T_C)$ if and only if $T^*$ and $T^*_C$ satisfy the following system of equations:

$$
\begin{align*}
X(T^*) - (W(K,T^*,T^*_C) - W(K-1,T^*,T^*_C)) &= 0 \\
X(T^*_C) - (W(K,T^*,T^*_C) - W(K-1,T^*,T^*_C)) &= 0
\end{align*}
$$

**Proof.** This proof is also covered in Appendix F, but we note that for a strategy $T_C$ in isolation, we may apply the proof of Theorem 3.3 with the initial assumption that our simple strategy simply ignores all majority candidates and selects only from the subset of minority candidates. With these two selection mechanisms in conjunction, however, the proof becomes a bit more complex.

Now, we would like to solve for the relationship between $T^*$ and $T^*_C$. We notice that Theorem 4.2 implies that $X(T^*) = X(T^*_C)$, and we already have a characterization of $X(T^*)$ such that:

$$
e^{-a} \cdot \left[ (1 - pq) + pq \cdot \frac{1}{j} (1 - e^{-a})^j \right] \leq X(e^{-a}) \leq e^{-a} (1 - pq) + pq \cdot \frac{1}{j} (1 - e^{-a})^j
$$

We can closely approximate $X(T^*_C)$ with $p \cdot q \cdot e^{-c}$. We begin by noting that this mechanism selects the best candidate only in the $p \cdot q$ scenario. Next, if we selected every minority contender, we would
simply have an \( e^{-c} \) probability of selecting the best candidate. However, we select minority contenders of Category C, which means that we must condition this probability on the fact that there have been majority contenders that have appeared before. Appendix B and experimentally extremely high values of \( e^{-c} \), however, allow us to argue that we are extremely likely to have seen a majority candidate better than best overall minority candidate in the \( p \cdot q \) case.\(^{12}\)

This is enough to characterize a relationship for a given \( p, q, \) and \( j \), and we choose a sample value in this range. To simplify the effect of \( j \) on this estimation, we choose the approximation:

\[
X(e^{-a}) = e^{-a} \cdot \left[ (1 - pq) + pq \cdot \frac{1}{j} \right]
\]

This is obviously larger than the lower bound, and whether \( e^{-a} \leq 1 - e^{-a}j \) depends of course on the values of \( e^{-a} \) and \( j \). However, experimentally, the values of \( e^{-a} \) that we examine are small enough that we expect this approximation to be reasonable. This leads to the characterization:

\[
P \cdot q \cdot e^{-c} = e^{-a} \cdot \left( 1 - pq \cdot \frac{j - 1}{j} \right)
\]

\[
e^c = e^a \frac{pq}{1 - pq \cdot \frac{j - 1}{j}}
\]

\[
c = a + \ln \left( \frac{pq}{1 - pq \cdot \frac{j - 1}{j}} \right)
\]

Now, due to the difficulty of characterizing \((W(K, T^*, T_C^*) - W(K - 1, T^*, T_C^*))\), we run a Bayesian optimization to estimate for the optimal \( T^* \) and \( T_C^* \). This optimization algorithm is suitable for our problem where our method of evaluating the expected success rate of a given \( T^* \) is simply to simulate it. Thus, we have a noisy and expensive black-box function and we seek a derivative-free optimization algorithm. Bayesian optimization proceeds by evaluating a small number of randomly selected function values and fitting a Gaussian process (GP) regression model to the results. The posterior distribution then dictates the next points to explore, updating with new information after each iteration. An acquisition function balances the exploitation of high expected-value areas, and a suitable alpha parameter on the Gaussian process allows us to adjust to the noisy estimations.\(^{20}\) We detail more about our this estimation in Appendix D.

We use these estimates for our standard scenario: \( p = .20, k = 7, \) and medium-bias to estimate \( T^* \) and \( T_C^* \) over \( q \in [0, 1] \). In order to smooth over the estimated results and to avoid over-fitting, we average over all of our predicted optimal points and impose a linear relationship between \( q \) and the optimal \( T^* \) value. We arrive at the equation:

\[
T^* = .0011q + .0034
\]

We set \( T_C^* \) accordingly. We graph this as a test of our estimation, comparing the simulated performance of our strategy to the optimal group-blind algorithm at every possible level of \( q \). It is trivial that the true optimal values should outperform the group-blind strategy at every \( q \) simply because we are optimizing over an additional variable. Thus we may lower bound \( W(K, T, T_C) \), of course, with at least the expected success rate of the optimal \( W(K, T) \).

\[
\max_{T, T_C} W(K, T, T_C) \geq \max_T W(K, T)
\]

\(^{12}\)For example, in the \( k = 7, p = .20, q = .20, j = 6 \) example that we examine in Section 6, we have \( e^{-c} = .887 \), which means that in the \( p \cdot q \) scenario, we are within 2.08e – 6 of the correct \( X(T) \).
And indeed, our window mechanism seems to outperform the optimal group-blind selection algorithm at every \( q \), an encouraging result. However, despite our ability to characterize what these strategies look like for a given \( q \), we must consider a practical problem: we are not likely to know exactly what \( j \) and what \( q \) are, and rather than lower bounding by choosing the least amount of bias that might exist – a strategy that likely means assuming no bias – we might want to be able to assess the performance of strategies over beliefs that are distributions over \( j \) and \( q \).

6 Algorithm Selection

We now put ourselves in the seats of someone implementing a dynamic selection algorithm in our model. “Group-blind algorithms are highly susceptible to low levels of implicit bias,” we might say, “but the affirmative action mechanisms are parameterized by a specific \( j \) and \( q \). How can I choose an alternative, and how do I know how these alternatives will perform?”

This is a reasonable concern. As the designer of the mechanism, it is reasonable to assume \( p, k \) are known to us.\(^{13}\) However, we now answer the question, “How might we propose alternative strategies and compare expected utilities when we have a distribution over values of \( q \) and \( j \)?”

We show that it is not only computationally feasible to evaluate alternative affirmative action mechanisms over arbitrary distributions of \( q \) and \( j \) but that it is simple. Then, we demonstrate this on our standard scenario with \( k = 7 \) and \( p = .20 \). We choose to calibrate the designated slot and window mechanisms on medium bias and a 20% chance of bias existing. We then compare these two strategies with the utopian group-blind strategy that assumes no bias, and evaluate the effectiveness of each strategy over a simple uniform distribution over \( j = [3, 6, 10] \).

\(^{13}\)We have not extended this problem to endogenizing a choice for \( k \) with a cost function. Thus, we must assume \( k \) is given exogenously.
6.1 Computational Feasibility

First, we note that no matter how complex the distribution of \( q \) is, the linearity of expectations makes computing the expected utility of a strategy simple. Note that \( q \) affects our strategies through two mechanisms: it can change how we set the parameters for our strategies, and it also changes how our strategies perform. In this scenario, however, we are comparing the performance strategies whose parameters have already been chosen. This means that \( q \) affects the utility of a strategy through the second pathway.

We calculate this, however, as a simple linear combination of the performance of the strategy when bias exists and when it does not exist. Thus, when we take the expectation of the utility of our strategy over a distribution of \( q \), the linearity of expectations allows us to recombine this into a simple function of the expected value of \( q \).

Thus, we only need to analyze the algorithm’s performance for two cases for each \( j \) – the case in which there is bias, and the case in which there is no bias.

Second, we show that for an arbitrarily good approximation of the utility \( \epsilon \), we may have an upper cutoff of bias \( \nu \) such that any probability that would’ve been assigned to \( j \geq \nu \) may be assigned to \( \nu \) instead. Thus for a given level of accuracy, we always have a finite support \( j \). To observe this for the window mechanism, we notice that our pre-calibrated strategy will always have the same pathway.

Consider a strategy evaluated on an arbitrary \( j_1 \) and \( j_2 \) such that \( j_2 > j_1 \). If we rewrite \( T \) and \( T_C \) as \( e^{-a} \) and \( e^{-c} \), then the \( c \) values will be defined by

\[
c(j_1) = a + \ln \left( \frac{pq}{1 - pq \cdot \frac{j_1 - 1}{j_1}} \right)
\]

\[
c(j_2) = a + \ln \left( \frac{pq}{1 - pq \cdot \frac{j_2 - 1}{j_2}} \right)
\]

We note, as we did in the proof of Theorem 3.3, that the strategies will only differ in behavior if a minority contender appears between in the gap between these exploration periods. We can see that as \( j \) increases, the exploration period becomes shorter, and this makes sense: the more bias there is, the less information major contender will give you on minority candidates. Thus, we are more likely to take Type C candidates. The probability that we select \( i \) minority contenders is given by substituting the appropriate values into Lemma 3.1:

\[
e^{c(j_2) - c(j_1)} \cdot \frac{(c(j_1) - c(j_2))^i}{i!}
\]

\[
= \frac{1}{i!} \cdot \exp \left( \ln \left( \frac{1 - pq \cdot \frac{j_1 - 1}{j_1}}{1 - pq \cdot \frac{j_2 - 1}{j_2}} \right) \right) \cdot \ln \left( \frac{1 - pq \cdot \frac{j_2 - 1}{j_2}}{1 - pq \cdot \frac{j_1 - 1}{j_1}} \right)^i
\]

\[
= \frac{1}{i!} \cdot \frac{1 - pq \cdot \frac{j_1 - 1}{j_1}}{1 - pq \cdot \frac{j_2 - 1}{j_2}} \cdot \ln \left( \frac{1 - pq \cdot \frac{j_2 - 1}{j_2}}{1 - pq \cdot \frac{j_1 - 1}{j_1}} \right)^i
\]

Now, as a result of \( j_2 > j_1 \), the term \( \frac{1 - pq \cdot \frac{j_1 - 1}{j_1}}{1 - pq \cdot \frac{j_2 - 1}{j_2}} \leq 1 \). In addition, because \( j_1, j_2 \in \mathbb{N} \), we know that this takes on its minimum value when \( j_2 \to \infty \), where we get a minimum value of \( \frac{1 - pq \cdot \frac{j_1 - 1}{j_1}}{1 - pq \cdot \frac{j_2 - 1}{j_2}} \). Now becomes clear that we can bound the distance of \( \frac{1 - pq \cdot \frac{j_1 - 1}{j_1}}{1 - pq \cdot \frac{j_2 - 1}{j_2}} \) from 1 by making \( j_1 \) larger, and because this function is continuous, \( \exists \nu \) such that this term is \( \epsilon \)-close to 1. We simply apply this logic to the probability of selecting 0 candidates, which is given by
\[ i = 0 \implies \frac{1}{i!} \cdot \frac{1 - pq \cdot \frac{j_1 - 1}{j_1}}{1 - pq \cdot \frac{j_2 - 1}{j_2}} \cdot \ln \left( \frac{1 - pq \cdot \frac{j_2 - 1}{j_2}}{1 - pq \cdot \frac{j_1 - 1}{j_1}} \right) = \frac{1 - pq \cdot \frac{j_1 - 1}{j_1}}{1 - pq \cdot \frac{j_2 - 1}{j_2}} \]

and since we know that we can make this term arbitrarily close to 1, we have that \( \exists \nu \) such that \( j_2 \geq j_1 \geq \nu \implies \) the probability of selecting 0 minority contenders between the two exploration periods is \( \epsilon \)-close to 1. Therefore, there is a \( 1 - \epsilon \) probability that the strategies won’t differ, in which case they have the same expected utility. When the strategies do differ, the difference in utilities in such a case belong to \([0,1]\), and therefore we have proven that expected utilities of the strategies differ by no more than \( \epsilon \) for all \( j \geq \nu \). Therefore, we argue that we can get an arbitrarily good approximation from a finite set.

Similarly, we could apply this analysis to the designated slot mechanism. We note that for our analysis of the designated slot mechanism, we choose \( T \) independently from \( j \), which means that the strategies differ only in the behavior of the designated slot and the size of that exploration period. We choose that size \( e^{-c} \) by optimizing \( c \) over the term:

\[ p \cdot q \cdot (1 - (1 - e^{-c})^{j-1}) \cdot e^{-c} \cdot c \]

We sketch out a proof: consider the two terms \( e^{-c} \cdot c \) and \( (1 - (1 - e^{-c})^{j-1}) \). The first is maximized at \( c = 1 \) and the second is maximized at \( c = 0 \) and is monotonically decreasing as \( c \) increases. Thus, our optimal \( c \) always falls in the range \([0,1]\). We construct lower and upper bounds.

\[ p \cdot q \cdot (1 - (1 - e^{-c})^{j-1}) \cdot \frac{1}{e} \leq \max_c p \cdot q \cdot (1 - (1 - e^{-c})^{j-1}) \cdot e^{-c} \cdot c \leq \max_c e^{-c} c = \frac{1}{e} \]

The lower and upper bounds converge as \( j \to \infty \), and thus for a given \( \epsilon_1 \) distance from the upper bound \( \frac{1}{e} \) we can find a \( j^* \) such that \( j \geq j^* \) implies that our function is within \( \epsilon_1 \) of \( \frac{1}{e} \). This allows us to compute a maximum distance we can be from \( c = 1 \), given that \( e^{-c} \cdot c \) is monotonic on the interval \([0,1]\). We can use this maximum distance as we did above to bound the difference between two bias levels \( j_1 \) and \( j_2 \) such that \( j_2 \geq j_1 \geq j^* \).

### 6.2 Analysis of Expected Utility

We analyze the performance of three strategies given a uniform distribution over \( j = [3, 6, 10] \).

---

\(^{14}\)The intuition may be described as such: as bias increases, the more \( T \) and \( C \) begin to become separate search processes on the majority and minority candidates, and the optimal \( C \) moves to what it would be on a single-candidate secretary problem, which is asymptotically the original solution: \( \frac{1}{e} \).
Comparison of Strategies
Medium Bias: j = 6, p = .20, k = 7

Success Rate

Probability of Bias (q)

*The error bars represent the 99% confidence intervals

Comparisons of Strategies
High Bias: j = 10, p = .20, k = 7

Success Rate

Probability of Bias (q)

*The error bars represent the 99% confidence intervals

Comparisons of Strategies
Low Bias: j = 3, p = .20, k = 7

Success Rate

Probability of Bias (q)

*The error bars represent the 99% confidence intervals

Figure 9.

Figure 10.

Figure 11.
Next, we show the expected utility over various values of $q$ given our uniformly distributed belief:

\[
\text{Expected Success with Uniform Distribution over } j = [3, 6, 10]
\]
\[\text{with } k = 7, \ p = .20\]

<table>
<thead>
<tr>
<th>$q$</th>
<th>Utopian Group-Blind</th>
<th>Designated Slot</th>
<th>Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.9436</td>
<td>0.9182</td>
<td>0.9429</td>
</tr>
<tr>
<td>0.10</td>
<td>0.9273</td>
<td>0.9094</td>
<td>0.9296</td>
</tr>
<tr>
<td>0.20</td>
<td>0.9110</td>
<td>0.9006</td>
<td>0.9163</td>
</tr>
<tr>
<td>0.30</td>
<td>0.8947</td>
<td>0.8919</td>
<td>0.9030</td>
</tr>
<tr>
<td>0.40</td>
<td>0.8784</td>
<td>0.8832</td>
<td>0.8898</td>
</tr>
<tr>
<td>0.50</td>
<td>0.8621</td>
<td>0.8744</td>
<td>0.8765</td>
</tr>
<tr>
<td>0.60</td>
<td>0.8458</td>
<td>0.8657</td>
<td>0.8632</td>
</tr>
<tr>
<td>0.70</td>
<td>0.8295</td>
<td>0.8569</td>
<td>0.8499</td>
</tr>
<tr>
<td>0.80</td>
<td>0.8132</td>
<td>0.8482</td>
<td>0.8366</td>
</tr>
<tr>
<td>0.90</td>
<td>0.7969</td>
<td>0.8394</td>
<td>0.8233</td>
</tr>
<tr>
<td>1.00</td>
<td>0.7806</td>
<td>0.8307</td>
<td>0.8101</td>
</tr>
</tbody>
</table>

Figure 12.

Encouragingly, we note that the window mechanism seems to improve expected utility at almost every $q$. This suggests that we might consider using implementing this mechanism. We also note that the designated slot mechanism outperforms the other algorithms at high levels of $q$.

### 6.3 Conditional Statistical Parity

We have focused on characterizing the scenarios in which affirmative action mechanisms can improve expected utility. However, we would be remiss not to analyze the disparity in error rates conditioned on the group label of the best candidate. We begin with the following lemma, relevant to the utopian group-blind algorithm that we have been examining.

**Lemma 6.1.** For the utopian group-blind algorithm with the assumption of no bias, the cost of implicit bias on the performance of the group-blind algorithm is entirely borne from an increased conditional error rate given that the best candidate is from the minority group.

This simply follows from our previous analysis. In the unbiased scenario, it is clear that the conditional error rates that are equal. However, when bias exists, it affects the performance of the utopian group-blind algorithm only when the best candidate is from the minority group.

This, combined with the steep effect of bias, is a concerning result from the perspective of conditional statistical parity.\(^{15}\) Thus, we show the disparity in conditional error rates for these strategies across low, medium, and high levels of bias and different probabilities $q$.

---

\(^{15}\) We borrow this definition from Sahil Verma and Julia Rubin: we have conditional statistical parity if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors $L$.\(^{23}\)
We see here that the designated slot mechanism affects conditional statistical parity drastically. This graph also allows us to understand that the window mechanism is replicating the behavior of the utopian group-blind algorithm at low $q$ because conditional statistical parity is affected far more modestly by this mechanism.
7 Conclusion

7.1 Discussion

In Section 3, we demonstrated the extremely high impact of “low” levels of implicit bias on the success rate of group blind algorithms. As a response, we proposed potential mechanisms to combat implicit bias in Section 4 and created bounds on their forms. In Section 5, we proved that the analysis of these was computationally simple even over complex beliefs and characterized the scenarios in which these affirmative action measures improved expected utility. Finally, we examined how implicit bias caused a large disparity in error rates between majority and minority groups.

The goal of our work was to improve our understanding of the effects of implicit bias on group-blind on a class of dynamic selection algorithms. In addition, we characterized a framework through which we can examine how diversity policies might improve not only expected fairness but expected utility. In particular, this work serves to nuance and extend the conversation about the role of protected classes in algorithms. A large body of legal literature has been written on anti-classification, where protected classes are not used to make decisions. This can be extended to orthogonalizing other predictors before dropping out the protected classes. However, in our work, we have showed that group-aware mechanisms can improve both expected utility and classification parity.\(^\text{16}\)

Perhaps more interesting than the question, “Do affirmative action policies work?” is the question, “When do they work?” Our characterizations in the dynamic selection process can provide some insight into these questions, and we attempt now to translate some of the analytic answers into intuitions.

For example, we might try to interpret these exploration periods as “certainty” levels. This is the intuition for the $X(T)$ analysis that powers the proof of Theorem 3.3: the marginal benefit of taking a contender must be example to the marginal cost. For a simple exploration period of 20\% of the candidates, for example, we could articulate, “We would be willing to use a slot on a candidate if we were at least 20\% sure they were the best candidate.” We could also understand the cost of implicit bias as a distortion of values underlying these beliefs — if 20\% of the population is minority and there is an 80\% of bias, a majority candidate who appears better than 20\% of the candidates that we have seen is no longer actually 20\% certain that they are the best candidate.

In addition, we can interpret our form of bias as the effects of having “favorites” or people might be preferred in any head-to-head direct comparison. If group effects caused five majority candidates to always be evaluated as more highly ranked than any minority candidate, we are already in our “medium-bias” scenario. This is stunning considering the large effects of bias in the medium bias scenario, and this speaks to the pervasive danger of having stereotypes about the groups that the best candidates belong to. For example, we characterize a scenario as having this kind of bias if the person doing the evaluation were to believe that the best CEOs are male.\(^\text{17}\)

Another important question is, “What kinds of policies work?” The window mechanism’s ability to improve on the naive group-blind algorithm across nearly all $q$ is striking, and it is encouraging that this mechanism improves fairness with respect to conditional statistical parity as well. In addition, with regards to fairness and performance in high-bias environments, the designated slot mechanism performs extremely well. However, we have only characterized only three strategies, and there are still far more to explore.\(^\text{18}\) Early analysis, for example, suggests that a strategy which sets a threshold for overall contenders from the majority class and then minority contenders is extremely bias-proof, albeit weaker at lower $q$. We have restricted our analysis to mechanisms which are implemented on top of or in conjunction with a simple group-blind strategy, but certainly we could begin searching over a broader class of strategies.

\(^{16}\)Classification parity is often used to denote the equality of common measures of predictive performance such as false positive and false negative rates across protected classes.

\(^{17}\)This is not an abstract concern: a recruiting algorithm developed by Amazon was scrapped because the deep-learning algorithm taught itself that male candidates were preferable and penalized resumes that included the word “women’s.”\(^\text{[22]}\)

\(^{18}\)We character the full separation strategy in Appendix C.
7.2 Further Exploration

One strong assumption in the model was the utility function – this is one of the many variants of the multi-choice secretary problem that has been traditionally studied, but an obvious extension is to consider other utility functions.\(^{19}\)

In addition, the centrality of \(q\) to this model suggests the importance of being able to assess the probability that bias indeed exists. Further research should be dedicated towards effective and simple methods of determining the probability that bias exists. For example, we might imagine examining dynamic strategies with shifting thresholds and using Bayesian inference on the order of candidates which we see in order to update our prior belief that bias exists. Alternatively, we might work on developing methods for estimating the probability that bias exists from data sets which show us complete or partial information from dynamic selection processes.

Finally, one of the most attractive parts of this model was the generality of the form of bias. As demonstrated in Section 2, nearly all other forms of bias could be transformed into our form of bias with no impact on the expected success rate or the error rates. However, this is not true with regards to the composition of the candidates that are selected: this might be an important issue to study.

In particular, the arrival order of the candidates with regards to group labels has the potential to drastically affect the diversity of the selected candidates and an exploration of its effects in the presence of uncertain bias has the potential to be fruitful. Early research seems to suggest that the arrival order of the candidates can have a drastic impact when there is bias.

8 Acknowledgements

I would like to thank my advisor, Professor Yang Cai for his patience and kindness in working with me to re-articulate this problem so many times. I would also like to thank Professor John Lafferty for suggesting using the Secretary Problem as a model for dynamic selection, and Professor Truman Bewley for his steadfast support and his thoughtfulness on questions of fairness and bias.

\(^{19}\)We should, however, consider the difficulty of analyzing such problems. As an example, we point to the 4\(^{th}\) secretary problem which remains an open problem – how to minimize rank with full information but no recall of preceding observations.\(^{21}\)
Appendix

A. Analysis of $f(T, C)$

We would like to calculate a lower bound for the probability that more than $k$ contenders appear after an exploration period of $T$ candidates but before the $T^{th}$ minority candidate appears. In order to do so, we prove that this probability is convex over a desired range and exploit Jensen’s inequality, which states:

If $X$ is a random variable and $\phi$ is a convex function, then $\phi(E[X]) \leq E[\phi(X)]$.

Now, consider the equation given by Lemma 3.1. We have that the probability of $k$ candidates appearing between $a$ and $b$ is given by:

$$p(k|a,b) = \binom{a}{k} \left( \frac{\log\left(\frac{b}{a}\right)}{k!} \right)$$

We take the second derivative with respect to $b$ and arrive at the following equation:

$$\frac{\delta^2}{\delta b^2} \left[ \frac{a}{b} \left( \frac{\log\left(\frac{b}{a}\right)}{k!} \right) \right] = \binom{a}{k} \cdot \frac{1}{b} \ln\left(\frac{b}{a}\right)^k$$

$$= \binom{a}{k} \cdot \frac{(\ln(b/a)^{(k-2)})(-3k \ln(b/a) + 2 \ln^2(b/a) + (k-1)k)}{b^3}$$

Now, we note that $\frac{b}{a} \geq 1 \implies \ln(\frac{b}{a}) \geq 0$. Thus, our second derivative is positive i.f.f

$$-3k \ln(b/a) + 2 \ln^2(b/a) + (k - 1)k \geq 0$$

For our analysis, we have a value for $a$ – this is the $T^*$ value that we choose as $T^* = e^{-k!}$. Because $b$ can take on any value from $a$ to 1, we have a maximum value of $\frac{b}{a}$ of $\frac{1}{e}$. In addition, for any given $k$ we can calculate the intervals for $\frac{b}{a}$ such that the function is convex over the entire interval – we simply note that the equation is a quadratic where the independent variable is $\ln(x)$.

As an example, when $a = 5! \frac{1}{e}$, $\frac{b}{a} \leq \frac{1}{e} \approx 0.3679$ and $\frac{b}{a} \approx 0.3679$. Separately, for $k = 5$, this function is convex over $\frac{b}{a} \in [1, 5.67]$. This, unfortunately, means that the $a$ that we have chosen does not guarantee convexity over the entire space of possible $b$ values. We can, however, guarantee convexity over $k = 7$, where the function is convex over $\frac{b}{a} \in [1, 14.76]$.

We compute the values that we can compute convexity over for the space of $k$ that we examine, and for $k \in [5, 10]$, we can guarantee convexity by solving for the probability that there are more than $k + 2$ candidates. For $k = 4$, it turns out we must jump to searching for $k \geq 7$ candidates to ensure convexity.

Now that have ensured convexity, we would like to compute the expected number of candidates seen before the exploration period for the designated slot ends. This can be calculated by taking the mean of the negative hyper-geometric distribution, which describes the number of majority candidates that we see before we see $C \cdot pN$ minority candidates. The parameters are accordingly set as $N, K = (1 - p)N$, and $r = C \cdot pN$. The mean of this distribution is $r \cdot \frac{N}{N - K + 1}$, and $\lim_{N \to \infty} r \cdot \frac{N}{N - K + 1} = C \cdot (1 - p)N$. Given that we will have seen $C \cdot (1 - p)N$ majority candidates and $C \cdot pN$ minority candidates, the expected number of total candidates seen when the exploration period for the designated slot mechanism ends is $CN$. 

29
B. Upper Bounding the Probability that the T selection mechanism affects the designated slot mechanism

We characterize the probability in the biased scenario that none of the \( z \) majority candidates who appear higher ranked than all minority candidates are chosen before a fraction \( C \) of the minority candidates are seen. First, it is trivial that we can ignore all non-\( j \) and non-minority candidates. Now, given a pool of \( N \) minority candidates and \( z \) candidates who cannot be drawn before we have seen \( c \) of the minority candidates, we construct this probability as:

\[
\begin{align*}
= & \prod_{i=0}^{CN-1} \left( 1 - \frac{z}{N-i} \right) \\
= & \prod_{i=0}^{CN-1} \left( \frac{N-z-i}{N-i} \right) \\
= & \left( \frac{N-z}{N} \right) \cdot \left( \frac{N-z-1}{N-1} \right) \cdot \ldots \cdot \left( \frac{(1-C)N-z}{(1-C)N} \right)
\end{align*}
\]

and now, cancelling out terms, we arrive at:

\[
= \prod_{i=0}^{z-1} \frac{(1-C)N-i}{i!} \cdot \prod_{i=0}^{z-1} \frac{N-i}{(1-C)N-i}
\]

and we therefore, we have that this limit for small \( z \) as \( N \to \infty \) is:

\[
\lim_{N \to \infty} \prod_{i=0}^{z-1} \frac{(1-C)N-i}{i!} \cdot \prod_{i=0}^{z-1} \frac{N-i}{(1-C)N-i} = (1-C)^j
\]

Thus, we have that an upper bound on the probability that the group-blind selection process selects a minority contender after a fraction \( C \) of the minority candidates have been seen.

C. Full Separation Between Groups

We might imagine, given the severe effect of bias, that we would want to examine the strategy with separates the two groups completely. This strategy has a major advantage: it is not affected by bias, because any shifts in the relative ranking of minority candidates to majority candidates won’t affect the intra-group rankings. A major cost of this separation, however, is that we throw away information that the two groups of candidates could tell us about each other. For example, we could avoid accepting a majority candidate who is not the best overall candidate if we used information from the minority group. However, studying these separation algorithms can give us a baseline for performance in high bias environments.

We denote the probability of winning with \( z \) candidates on a pool of size \( pn \) as \( w(z|pn) \). We can think of this algorithm as running two separated multi-choice secretary problems, and we thus solve the simple optimization problem:

\[
\max_{z \in \mathbb{Z}_{\geq 0}} \max_{z \leq k} \left( p \cdot w(z | pn) + (1-p) \cdot w(k-z | (1-p)n) \right)
\]

We set the exploration periods on each process appropriately for the given number of candidates, assuming that \( N \to \infty \). These are \((z!)^{\frac{1}{z}}\) and \((k-z)!(\frac{k}{z})^{\frac{k}{z}}\) respectively. Finally, we rewrite these win probabilities with Theorem 3.2.
\[
\max_{z \leq k} p \cdot e^{(z!)/z!} \sum_{i=1}^{z} \frac{(z!)^{1/2}}{i!} + (1 - p) \cdot e^{(k-z)!/(k-z)!} \sum_{i=1}^{k-z} \frac{(z!)^{1/2}}{i!}
\]

We compare the win probability of these strategies to the upper-bound for group-blind algorithms in low, medium, and high bias environments.

This is an encouraging result: at around \( p = .20 \), we have that the full-separation algorithm begins to outperform group-blind algorithms at these levels of bias. However, when we fix \( p = .20 \) and factor in the probability that bias exists, the results are substantially less encouraging.
Thus, we have omitted this strategies from the main body of our analysis.

D. Simulation Information

For all data that contains standard error bars, we run each simulation 100,000 times and calculate our standard error accordingly. In order to approximate $N \to \infty$, we use $N = 10,000$. The code is provided in a with the project.

For the Bayesian optimization, we use the implementation:

```
https://github.com/fmfn/BayesianOptimization
```

We set the alpha parameter equal to $2.5e - 3$, and run 20 iterations with 2 random points and 3 specified “probe” points.

E. Applying Theorem 3.3 to $W(K - 1, 1, T, C)$

Consider our designated slot strategy, where we select any minority contender that $T$ does not select after an exploration period of $C$ minority candidates. We would like to show the following:

**Theorem.** $T^*$ and $C^*$ maximize $W(K - 1, 1, T, C)$ if $T^*$ and $C^*$ solve:

$$X(T^*) - (W(K - 1, 1, T^*, C^*) - W(K - 2, 1, T^*, C^*)) = 0$$

$$X(C^*) - (W(K - 1, 1, T^*, C^*) - W(K - 2, 1, T^*, C^*)) = 0$$

**Proof for $T_C$.**

First, we note that $W(K - 1, 1, T, C)$ is defined over $K \in \mathbb{N}$ and $T, C \in [0, 1] \cap \mathbb{Q}$. Thus, our partial derivative with respect to $C$ is defined over the rationals, and we express $C$ as $\frac{a}{b}$ and $h$ as $\frac{c}{d}$.

$$\frac{\delta}{\delta C} W(K - 1, 1, T, C) = \lim_{h \to 0} \frac{W(K - 1, 1, T, C - h) - W(K - 1, 1, T, C)}{h}$$

$$= \lim_{\frac{h}{d} \to 0} \frac{W(K - 1, 1, T, \frac{ad - cb}{bd}) - W(K - 1, 1, T, \frac{ad}{bd})}{\frac{d}{d}}$$

$$= \lim_{\frac{a}{d} \to 0} \frac{d}{c} \cdot \left( W(K - 1, 1, T, \frac{ad - cb}{bd}) - W(K - 1, 1, T, \frac{ad}{bd}) \right)$$

And we can again directly compare the performance of these two terms by comparing the probability that there are minority contenders that appear between them. Notice, however, that because we have two strategies running simultaneously, there is always the possibility that a minority contender appears that the $T$ selection mechanism would’ve selected. In this case, the strategies would not differ, because they have the same $T$. Thus, we denote the probability that the $T$ strategy would not have selected this candidate with $\tau$. $\tau$ is a function which represents the probability that the $T$ strategy would not pick the minority contender, and it is a function of $p, q, k, j$, and the group of candidates seen. We use $\tau(q, p, k, j, \frac{a}{b})$ to represent the probability that a candidate seen at $T_C = \frac{a}{b}$ would be selected by $T$. We use $\tau(q, p, k, j, \frac{a}{b}, \frac{c}{d})$ to represent the probability that a candidate seen anywhere in between $\frac{a}{b} - \frac{c}{d}$ and $\frac{a}{b}$ would be selected by $T$.

For the ease of notation, we will first denote this function simply with $\tau$. Now, let us denote the probability of selecting a contender that would not be selected by $T$ with $p(i, a, b) \cdot \tau = p_r(i, a, b)$. There is only one
slot, so we either select one contender or none in the minority candidates between these two explanation periods.

Now, we must simply prove that the probability of selecting exactly one candidate does not also go to 0. We argue that \( \tau(q, p, k, j, \frac{a}{b} - \frac{c}{d}, \frac{a}{b}) \) is continuous, so we have that \( \lim_{\frac{c}{d} \to 0} \tau(q, p, k, j, \frac{a}{b} - \frac{c}{d}, \frac{a}{b}) \) converges to a term \( \tau(q, p, k, j, \frac{a}{b}) \).

Therefore, for the probability of selecting 1 candidate, we have that:

\[
p_\tau(1, \frac{a}{b} - \frac{c}{d}, \frac{a}{b}) = \lim_{\frac{c}{d} \to 0} \tau(q, p, k, j, \frac{a}{b} - \frac{c}{d}, \frac{a}{b}) \cdot p(1, \frac{a}{b} - \frac{c}{d}, \frac{a}{b})
\]

\[
= \tau(q, p, k, j, \frac{a}{b}) \cdot \lim_{\frac{c}{d} \to 0} \frac{d}{c} \cdot \left(1 - \frac{cb}{ad}\right) \cdot \ln \left(\frac{1}{1 - \frac{cb}{ad}}\right)
\]

\[
= \tau(q, p, k, j, \frac{a}{b}) \cdot \frac{b}{a}
\]

Now, we make use of the fact that \( p_\tau(i, a, b) \) is a probability distribution to observe the following.

\[
p(0, ad - cb, ad) + p(1, ad - cb, ad) = 1
\]

We can use this to compute the probability that \( i = 0 \) in our derivative, and thus, canceling out terms and proceeding with the proof from Theorem 3.3, we have that the partial derivative is equal to:

\[
\frac{\delta}{\delta T_C} W(K, T, T_C) = \tau(q, p, k, j, \frac{a}{b}) \cdot \frac{b}{a} \left(X(\frac{a}{b}) + W(K - 1, 0, T, \frac{a}{b}) - W(K - 1, 0, T, \frac{a}{b})\right)
\]

Finally, we argue that \( \tau(q, p, k, j, \frac{a}{b}) > 0 \) for all non-degenerate cases. The probability that a group-blind algorithm does not select a minority contender is always non-zero. Thus, by the zero product property, we have that:

\[
\tau(q, p, k, j, \frac{a}{b}) \cdot \frac{b}{a} \left(X(\frac{a}{b}) + W(K - 1, 0, T, \frac{a}{b}) - W(K - 1, 0, T, \frac{a}{b})\right) = 0 \implies \left(X(\frac{a}{b}) + W(K - 1, 0, T, \frac{a}{b}) - W(K - 1, 0, T, \frac{a}{b})\right) = 0
\]

Proof for T.

We note that the proof for \( T \) is nearly exactly the same as for \( C \): we simply create the random variable which denotes the expected probability that the selection mechanism \( C \) would not have taken the candidate and work with the random variable in the same way.

Appendix F. Applying Theorem 3.3 to \( W(K, T, T_C) \)

Consider the strategy \( W(K, T, T_C) \), where \( T_C \) selects minority candidates of Category \( C \) after an exploration period of \( T_C \) minority candidates. We would like to show that Theorem 3.3 applies to our strategy as follows:

**Theorem.** \( T \) and \( T_C^* \) maximize \( W(K, T, T_C) \) if \( T \) and \( T_C^* \) solve:

\[
X(T) - (W(K, T, T_C) - W(K - 1, T, T_C^*)) = 0
\]
\[
X(T_C^*) - (W(K, T, T_C) - W(K - 1, T, T_C^*)) = 0
\]

\[20\]We remind the reader that Category \( C \) represents minority candidates better than all minority candidates already seen, but not better than all candidates overall.
Proof for $T_C$.

First, we note that $W(K, T, T_C)$ is defined over $K \in \mathbb{N}$ and $T, T_C \in [0, 1] \cap \mathbb{Q}$. Thus, our partial derivative with respect to $T_C$ is defined over the rationals, and we express $T_C$ as $\frac{a}{b}$ and $h$ as $\frac{c}{d}$.

$$\frac{\delta}{\delta T_C} W(K, T, T_C) = \lim_{h \to 0} \frac{W(K, T, T_C - h) - W(K, T, T_C)}{h}$$

$$= \lim_{\frac{c}{d} \to 0} \frac{h}{d} \left( W(K, T, \frac{a - cb}{bd}) - W(K, T, \frac{a}{b}) \right)$$

And we can again directly compare the performance of these two terms by comparing the probability that there are minority contenders that appear between them. Even though we have two strategies running simultaneously, there is no overlap in the candidates that they might select. However, we do have to nuance the analysis by noting that we only select the minority contender if we have seen a majority candidate before that appears more highly ranked. Let us denote this probability with $\nu$.

$\nu$ is a function which represents the probability that this candidate is of Category C, and it is a function of $p, q, k, j$, and the group of candidates seen. We use $\nu(q, p, k, j, \frac{a}{b})$ to represent the probability that a candidate seen at $T_C = \frac{a}{b}$ is of Category C. We use $\nu(q, p, k, j, \frac{a}{b} - \frac{c}{d}, \frac{a}{b})$ to represent the probability that a candidate seen anywhere in between $\frac{a}{b} - \frac{c}{d}$ and $\frac{a}{b}$ is of Category C.

For the ease of notation, we will first denote this function simply with $\nu$. Now, let us denote the probability of selecting a contender that would not be selected by $T$ with $p(i, a, b) \cdot \nu = p_\nu(i, a, b)$. We note that this is always less than the probability of simply selecting a contender, so we can write that:

$$p_\nu(i, a, b) \leq p(i, a, b)$$

This implies that because we showed in Theorem 3.3 that $p(i, a, b)$ converges to 0 for all $i \geq 1$, $p_\nu(i, a, b)$ must do so as well. Now, we must simply prove that the probability of selecting exactly one candidate does not also go to 0. We argue that $\nu(q, p, k, j, \frac{a}{b} - \frac{c}{d}, \frac{a}{b})$ is continuous, so we have that $\lim_{\frac{c}{d} \to 0} \nu(q, p, k, j, \frac{a}{b} - \frac{c}{d}, \frac{a}{b})$ converges to a term $\nu(q, p, k, j, \frac{a}{b})$.

Therefore, for the probability of selecting 1 candidate, we have that:

$$p_\nu(1, \frac{a}{b} - \frac{c}{d}, \frac{a}{b}) = \lim_{\frac{c}{d} \to 0} \nu(q, p, k, j, \frac{a}{b} - \frac{c}{d}, \frac{a}{b}) \cdot p(1, \frac{a}{b} - \frac{c}{d}, \frac{a}{b})$$

$$= \tau(q, p, k, j, \frac{a}{b}) \cdot \lim_{\frac{c}{d} \to 0} \frac{d}{c} \left( 1 - \frac{cb}{ad} \right) \cdot \ln \left( \frac{1}{1 - \frac{cb}{ad}} \right)$$

$$= \nu(q, p, k, j, \frac{a}{b}) \cdot \frac{b}{a}$$

Now, we make use of the fact that $p_\nu(i, a, b)$ is a probability distribution to observe the following.

$$\sum_{i=0}^{\infty} p(i, ad - cb, ad) = 1$$

This implies that:

$$\lim_{\frac{c}{d} \to 0} \left( \sum_{i=0}^{\infty} p_\nu(i, ad - cb, ad) \right) = \lim_{\frac{c}{d} \to 0} \left( \frac{d}{c} \right)$$
We can use this to compute the probability that \( i = 0 \) in our derivative by subtracting out the probability that there are 1 or more contenders – which is simply the probability that there is 1 contender.

\[
i = 0 \implies \lim_{\frac{d}{c} \to 0} \left( \frac{d}{c} - \sum_{i=1}^{\infty} \text{Terms} \right) = \lim_{\frac{d}{c} \to 0} \frac{d}{c} - \sum_{i=1}^{\infty} \lim_{\frac{d}{c} \to 0} \text{Terms} = \lim_{\frac{d}{c} \to 0} \left( \frac{d}{c} - \nu(q, p, k, j, a \cdot b) \cdot \frac{b}{a} \right)
\]

Thus, canceling out terms and proceeding with the proof from Theorem 3.3, we have that the partial derivative is equal to:

\[
\frac{\delta}{\delta T} W(K, T, T_C) = \nu(q, p, k, j, a \cdot b) \cdot \frac{b}{a} \left( X(\frac{a}{b}) + W(K - 1, T, \frac{a}{b}) - W(K, \frac{a}{b}) \right)
\]

Finally, we argue that \( \nu(q, p, k, j, a \cdot b) > 0 \) for all non-degenerate cases. The probability that a group-blind algorithm does not select a minority contender is always non-zero. Thus, by the zero product property, we have that:

\[
\nu(q, p, k, j, a \cdot b) \cdot \frac{b}{a} \left( X(\frac{a}{b}) + W(K - 1, T, \frac{a}{b}) - W(K, \frac{a}{b}) \right) = 0 \implies \left( X(\frac{a}{b}) + W(K - 1, T, \frac{a}{b}) - W(K, \frac{a}{b}) \right) = 0
\]

**Proof for \( T \).**

This proof simply follows Theorem 3.3 exactly, noting that the addition of the \( T_C \) mechanism does not select any of the same candidates, and therefore does not affect our analysis.
References


