Field of Algorithmic Fairness:

I hope that this work will be a part of the growing body of literature of algorithmic fairness. Data collection and algorithmic prediction has become increasingly prevalent in our social and economic lives. From credit scores to hiring tools\(^1\), the consequences of implicit bias in algorithms are higher than ever. As a result, a growing number of groups have been dedicated towards improved fairness and interpretability in both shallow and deep learning algorithms.

Within these groups, there are several growing fields of research: training algorithms with mechanisms to prevent disparate outcomes, plugins to add interpretability to algorithms so that bias is more transparent, and studying existing algorithms to find bias.\(^2\) However, questions of fairness are complex and multi-disciplinary. Many researchers, such as Sam Corbett-Davies from Stanford and Robert Williamson have conceived of the “cost of fairness” as the loss imposed in moving from the unconstrained optimization problem to the constrained problem.\(^3\)

However, I aim to examine a branch of algorithmic fairness in which both algorithm performance and classical measures of fairness can be improved – economic situations in which actors have *implicit bias*, and ways to make algorithms robust against this. Specifically, I will analyze the k-choice secretary problem as a model for implicit bias in dynamic hiring decisions. This work is partially inspired by Jon Kleinberg’s paper, *Selection Problems in the Presence of Implicit Bias*, which deals with the static hiring problem, and by a comment by Professor John Lafferty made when I presented the paper to him.

The K-Choice Secretary Problem:

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. It has been studied extensively in literature, 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.

\(^1\) Dastin, Jeffrey. “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women.” Reuters.
\(^2\) For example, one recent study found that natural language processing algorithms can encode gender biases as simple as associating the word nurse more closely with the word she than with the word he.
\(^3\) Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. “Algorithmic Decision Making and the Cost of Fairness.” Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1:797-806.
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 "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 (Glasser, Holzager, and Barron 1983). 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 \( k \) 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.

**Abstract:**

Experimental studies have shown that implicit bias is an important driver of discriminatory and unfair outcomes in selection-based activities, from hiring decisions to school admissions. We propose a theoretical model for studying the effects of implicit bias on optimal dynamic selection, where a decision must be reached about a candidate immediately after interviewing them, and propose a procedural remedy for detecting the presence of implicit bias and remediying it. We use the multi-choice secretary problem as a model for a manager trying to move candidates to the final round of interviews – the shortlist for the job. The success of the manager is determined by whether this shortlist contains the best candidate. However, we extend this model by making two assumptions. The first is that there are two groups within this hiring pool: one majority group and one minority group. These are identifiable to the manager. The second is that the minority group may or may not be subject to a minimum amount of bias, and the manager has a belief about this probability. I will attempt to show the optimal algorithm in such a scenario, and empirically demonstrate its improvement over the no-bias optimal algorithm. By employing a mechanism similar to the Rooney Rule\(^4\) – we improve both the expected diversity of the candidates selected and the expected utility.

---

\(^4\) 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.
Abstract II:

In the same theoretical set-up, we attempt to analyze the effects of ordering on the performance of such an algorithm. The original secretary problem deems that you take the candidates in a random order. However, in the extension of the problem, given that quality of the group is observable, how does the ordering in which you observe candidates with a given algorithm affect the outcome in terms of success and diversity?

Deliverables:

I aim to produce a theory-driven thesis that contains my original work on the problem that I have posed below. I have two different abstracts that I am considering work on – the second more tractable than the first, and I aim to work on both simultaneously.

In the end, I expect to have 10 – 15 pages of original work, with a few theorems proved by the end of my work and incorporated into a larger research paper. In addition, I will write simulations in Python to check the performance of my algorithm and provide the code and results.

Preliminary Ideas for Theorems:

1) Go over and re-derive the optimal strategy for the k-choice secretary problem when there is no bias
2) Identify the optimal strategy when we are sure there is bias, and we are aware of the minimum amount but have no information about the distribution
3) Show how Bayesian updating can be performed on the sequence of candidates that we have seen so far. This may be tricky, because we keep the form of the bias as general as possible.
4) Provide the optimal strategy for our scenario, a manager with a prior about the chance of having implicit bias.

Nature of Meetings

I will meet with Professor Cai Yang (on average) once every two weeks. Because of the nature of the theoretical work, I will generally try to meet only when I have made progress on the work. This, however, does not only include original results – understanding the proofs behind optimal algorithms, for example, in different iterations of the k-choice secretary problem is essential.

In addition, I will construct some empirical analyses of the effectiveness of my algorithm. This will serve to both confirm the correctness of my mathematical proofs and to provide a path toward more steady progress (as opposed to the choppy nature of progress in theoretical work).
Possible References