Recruiting in the White House: Addressing Fairness in Automated Hiring Algorithms

Jason D. Lazar
Department of Computer Science
Stanford University
jasonl24@stanford.edu

Proud B. Mpala
Department of Computer Science
pmpala@stanford.edu

Eva I. Prakash
Department of Computer Science
eprakash@stanford.edu

Josh Singh
Department of Computer Science
jsingh5@stanford.edu

Video: https://tinyurl.com/wey2srus

1 Introduction

1.1 Our Motivation

It is evident that the job application process is only becoming more competitive for both entry-level careers and experienced professional roles. As applications continue to increase, automation is becoming an integral part of the recruitment process. It is well established that automated systems are currently in place to auto-screen applicant resumes and listed qualifications. Systems like Applicant Tracking Software (ATS) are implemented to save recruiters time and effort as they work to narrow down potential candidates. However, we predict that as artificial intelligence (AI) and machine learning models are increasing in capability, such systems will begin to automate HR-related components to job applications, like initial behavioral interviews.

More specifically, we envision a future where companies will rely on trained models to conduct virtual interviews by using both computer vision and NLP techniques. It is likely that more decision making power will be entrusted to artificial models, saving companies time, effort, and money. Thus, when selecting candidates, it is imperative to address fairness concerns to ensure an equitable and impartial hiring process, as most companies promise to the public.

This paper focuses on investigating the fairness implications of using automated hiring algorithms in the context of White House recruitment. By analyzing the application of AI to conduct interviews, we aim to address individual and group fairness concerns. Through a mathematical framework, we asses the extent to which these algorithms adhere to fairness principles and propose strategies to mitigate biases and promote equitable hiring outcomes. This research contributes to the broader conversation on algorithmic fairness in recruitment processes, with implications for improving diversity and representation in political institutions.

1.2 Why We Choose the Context of White House Recruiting

There are many contexts in which fair hiring applies, but we focus on the White House, where the stakes of fair hiring are particularly high. The White House is one of the most powerful institutions in the world and makes decisions that affect billions of people, so it is especially crucial to hire fairly to ensure that all those relevant to White House politics are accurately represented. In general, individuals are often biased toward issues that are directly relevant to or affect them, so when there is a lack of diversity in governing bodies, there is a significant risk of not addressing problems that affect a wide range of citizens. As it stands, the White House hiring process exhibits biases favoring applicants

already connected to current staffers, which potentially exacerbates homogeneity of thought and potentially overlooks well-qualified candidates. Integrating AI into hiring has the potential to help these systems, but it needs robust guardrails, which have often been disregarded in the past. A large number of publications discuss potential issues with using AI, including semantic bias, cultural bias, and socioeconomic bias, to name a few. These concerns arose when the government passed the AI Bill of Rights, which includes Algorithmic Discrimination Protections. (The United States Government, 2022a) To mitigate the issues with regular and AI-assisted hiring in the high stakes environment of the White House, we construct a model that aims to hire a diverse set of individuals, with varying political identities, thoughts, backgrounds, and identities, but all with enough merit and qualification. (The United States Government, 2022b).

Inspired by the context of White House recruiting, in this paper, we present a mathematical framework to address the complex notions of fairness that would aid in the hiring process of the White House. We hope this model will harness the strengths that exist in the current system, but make the overall process more fair.

2 Definitions

2.1 Context

We will be considering a setting where we have job candidates who initially indicate their department preference rankings, among other features to the White House. The White House has decided to leverage an LLM powered conversational virtual agent to conduct the first behavioral interview for this pool of candidates. During an interview for a given department, the agent asks a series of adaptive questions and then generates a word-for-word transcript of the candidate's interview. Then another virtual agent takes this report along with candidate preferences and information on the hiring quotas in the White House and makes a decision about which department to hire the candidate to or not hire the candidate at all.

2.2 Formal Model

We model our job applicant population P as individuals from some universe $\mathcal{X}$ so $P \subset \mathcal{X}$. The population includes disjoint groups $G_1, G_2, \ldots G_m$ where $G_i \subset \mathcal{X}$. The institution that the candidates are applying to consist of k departments $d_1, d_2, \ldots, d_k$. Candidate i's preference ranking can be denoted as an ordered list of departments they want to apply to in order of decreasing priority $r_i = d_1, d_2, d_3$ implies the candidate prefers department 1, 2, 3 (here we assume that the White House permits applying to a maximum of 3 departments but the list can have arbitrary departments in the general case).

We also define T as a space containing all possible interview transcripts. Applicants interact with a deterministic automated interview algorithm, A, that produces a transcript of the interview $A : \mathcal{X} \to T$. A asks questions and followup questions somewhat based on applicant responses and provides a full word-for-word transcript as its output. We then have a deterministic decision algorithm informed by the White House, D, that takes the transcript and decides which department the candidate will be accepted to or $\emptyset$ if a student is not accepted to any department: $D : T \times \mathcal{X} \to [k] \cup \emptyset$, k is the number of departments that are present in the institution that our candidates are applying to. The interview algorithm A and the decision algorithm D have no knowledge of each other we treat them as separate entities otherwise we risk a case where there might be collusion. The choice for a deterministic setting is to simplify the model but notions of fairness that will be raised later can be extended to a randomized settings.

We model our institution to have quotas for each group. Q is the total number of candidates that can be hired. We define Q_i to be the number of spots in total that are reserved for candidates from group i. We also make the assumption that $|P| >> |Q|$ so that we do not have to worry about the case where fewer than needed applicants applied.

3 Fairness Concerns

In this section, we develop various abstractions that we can use to scrutinize the output of our hiring algorithms. We draw inspiration from the college admissions example that was discussed in class

where similar individuals were accepted with similar probability for individual fairness. Unlike that situation, we make the argument that Preference Informed Individual Fairness (PIIF) best captures the fairness concerns that may arise with respect to individual applicants. As will be captured by the definitions below the main idea is that for individual fairness, two individual that are similar should be treated similarly in terms of their outcomes. As for group fairness, we draw upon the US law notions in class, the San Francisco vs New York question in PSet 3, and other concepts to discuss quotas and decisions independent of protected attributes.

3.1 Individual Fairness

The hiring system operates in two stages, first the transcription stage A, then the decision stage carried out by D.

For metric fairness, two similar individuals $x, y \in \mathcal{X}$ when they interact with the LLM algorithm A should have a similar transcript. To capture the similarity between two individuals, we introduce a distance metric $d(x, y)$. We also define a metric that captures the sentiment score of each individual's transcript: $s : T \to [0, 1]$.

Thus, our algorithm A satisfies metric fairness if:

$$|s(T_x) - s(T_y)| \leq d(x, y) \tag{1}$$

We also need to ensure that the decision algorithm satisfies metric fairness. Traditionally we can use the sentiment distance $|s(T_x) - s(T_y)|$ as the new distance metric in which case our decision algorithm D satisfies metric fairness if $D_{TV}(D(T_x), D(T_y)) \leq |s(T_x) - s(T_y)|$ but if we do this we run into the issue that the decision algorithm D does not take into account the preferences of the two applicants. For this reason, we turn to PIIF which is suitable in cases where we want to enforce metric fairness as well as take into account the preferences of the applicants (Kim et al., 2019). Inspired by PIIF, we define a new distance metric $d_T((r_x, T_x), (r_y, T_y))$ which takes in the preferences of two individuals along with their transcripts and indicates how similar these two individuals are given their transcripts and preferences. To enforce similarity in distribution of outcomes given our algorithm D we enforce that:

$$D_{TV}(D(x, T_x), D(y, T_y)) \leq d_T((r_x, T_x), (r_y, T_y)) \tag{2}$$

From the two conditions (1) and (2) that enforce metric fairness with respect to algorithm A and D, we ensure that the hiring process is fair with respect to each individual.

3.2 Group Fairness

In our model, we make the assumption that the institution that is hiring has thoroughly researched and came up with a way to determine the the quotas for each group. We want the total number of applicants that are admitted to be equal to the total number of job openings. Moreover, for each group, we want the number of applicants hired from that group to be proportional to the spots reserved for individuals from that group. To ensure group fairness, our decision algorithm D should align with the set quotas as follows:

$$E[\textstyle\sum_{j=1}^{m} \mathbf{1}[D(j, T_j) \neq \emptyset]] = Q$$
$$E[\textstyle\sum_{j=1}^{m} \mathbf{1}[D(j, T_j) = j]] = Q_j$$

where the total $Q = \sum_{i=1}^{k} Q_k$.

Getting back to our algorithm A in the context of group fairness, a study that investigated a similar system noted that such LLM systems tend to be very sensitive to some attributes like political affiliation and pregnancy status. (Veldanda et al., n.d.) This study along with the USA government's Act on hiring gives us an angle to approach group fairness in this context. (US-, n.d.). Also, in class, we discussed fairness through blindness, and we hope to get closer to this ideal by taking protected classes into account. To formalize this, we introduce protected attributes $a_1, a_2, a_3, \ldots, a_n$. Our transcription algorithm A should not discriminate based on these attributes. If a particular group has a heavy presence of a protected attribute then even if that attribute is not reported, the LLM can use

other attributes as a proxy and still discriminate based on a protected attribute even if the protected attribute was not reported.(Alvero et al., 2020) For this reason we want that our transcript algorithm to produce a transcript independent of protected attribute.

We claim that this issue can be captured by the distance metric $d(x, y)$ which appears in (1). Specifically, if our distance metric is blind to protected attributes, then it will enforce group fairness in relation to the particular attribute. So we define attribute blind distance metric. We say that a distance metric is blind to a particular attribute a if:

$$d(x_{\text{with attribute } a}, x_{\text{attribute } a}) = 0$$

i.e the distance metric does not see a difference when a particular attribute is marked present or absent in a particular individual x.

4 Related Work

To understand more about how to make application selection process fairer, we draw upon research primarily focusing on college admissions and somewhat related to similar fields such as income or occupation prediction. Most of the research we analyze has to do with college admissions, as college admissions and job applicant selection are similar in that both processes are holistic, take into account a wide range of public and protected information, have a significant impact on applicants' lives, and are open-ended. The work that has made these predictors and processes more fair could have relevant applications to our model.

Alvero et al. attempted to explore the fairness of college admissions by collecting a dataset of approximately 280K applicant essays for selective public universities and training a multinomial naive Bayes model, a logistic regression model, and a deep neural network to predict reported household income (RHI) and reported gender (RG) from the essays. All models achieved approximately 70% accuracy on RHI and approximately 80% accuracy on RG, which indicate how specific words or combinations of words can be used to infer sensitive attributes. From the naive Bayes model, the frequency ratios $FR(w) = \frac{p(w|0)}{p(w|1)}$ were extracted, indicating how frequently each word w appeared in writing with label 0 relative to writing with label 1. Such experiments found that words like "hardware," "LEGOS," and "chess" were indicators for boys, and words like "scouting", "rowing," and "Switzerland" were indicators of RHIs above the nationwide median. This work aimed to emphasize the importance of balancing how AI researchers are often more "concerned with fairness and bias at the population level," whereas "admissions officers tend to emphasize fairness of evaluation for individual applicants" (Alvero et al., 2020). This research is important as it underscores the importance of having unbiased agents that produce transcripts and decisions blind to protected attributes.

Agarwal et al. experimented with "binary classification subject to fairness constraints with respect to pre-defined protected attributes" (e.g. race and sex) in a variety of contexts, including adult income prediction over $50K per year, recidivism rates, prediction of prestigious occupation, and, most relevant for our purposes, law school admissions (i.e. passing the bar exam). The algorithmic setup enforced either demographic parity and equalized odds on the classifier loss function in order to create a cost-sensitive classification problem with the "lowest (empirical) error subject to the desired constraints." Across all datasets, the approach was able to show promise in reducing disparity with respect to various protected attributes while maintaining the overall accuracy of the binary classifier. This work provided helpful insight into the optimization of the tradeoff between accuracy and "any (single) definition of fairness given access to protected attributes" (Agarwal et al., 2018).

More recently, Zehlike et al. worked to formalize the algorithmic construction and fairness concerns of college admissions. They described the admissions process as prioritizing likelihood of success in college and strong interest in various majors as well as "forming a demographically diverse group both overall and in each major." To model this setup, they conceptualized a score-based ranker pipeline on a dataset of college applicants, involving the consolidation of demographic attributes (e.g. race and sex), numerical attributes (e.g. GPA, SAT, ACT), and vector representations extracted from the applicants' essays into a formula given by admissions officers that returns the top-K highest-scoring applicants in ranked order to be interviewed and potentially admitted. To enforce measures of fairness, the authors calculated metrics such as normalized discounted difference, normalized discounted ratio, and KL-divergence between various demographic groups and the overall population. This

work was useful in forming the mathematical representations around admissions processes and their corresponding fairness concerns (Zehlike et al., 2022).

5 Limitations and Next Steps

While we believe this model effectively contributes to making the White House hiring process more fair, there are still some limitations of our model and next steps to take to ensure our model is more robust.

One assumption our model relies upon is that the agents are deterministic and will give the same output given the same input. However, LLM agents can exhibit non-perfect and non-predictable behavior; even given the same candidate, it can output two different transcripts by asking different follow-up questions and make two different hiring decisions. Another assumption we made about LLMs is that they are independent of each other. To be more concrete, we noted that agents A and D have no knowledge of each other. While it may be the case that A and D have no knowledge of each other, as many LLMs are trained on extensive datasets with possible overlap, there can still be collusion.

Another limitation is that only two features are taken into account for Agent D: the transcript from Agent A and the applicant's preferences. (It is important to note that other features can be embedded into the transcript.) As Agent A is created by the White House, we hope it will output sufficient information in the transcript for Agent D to make a decision. However, we may have scenarios where we would want Agent D to take into account more information. For example, we might want to take into account the applicant's official college transcripts and letters of recommendation or information about the relevant department's funding and headcount which cannot be captured in the transcript or preferences.

We also have the limitation that we only take three preferences of the applicants. Applicants may want to work in many departments, some of which may have more space than others. We limited the number of preferences to three as that is how the current White House job application system works, but it might be better for the White House and the applicants if they can rank more departments as it only increases the chances of a hiring occurring.

To improve this work, there are a couple of potential future steps. We could try to incorporate more information from the department and its needs into the hiring decisions. It might be worthwhile to understand how to decrease collusion between the agents through strategies like fine-tuning. One major next step could be to think about how to add more independent LLM agents together to reduce the chances of collusion and hallucination. Also, these LLMs could specifically cross-check each other's work in specific facets so group and individual fairness metrics are upheld. We may also consider incorporating normalized discounted difference, normalized discounted ratio, and KL divergence like the researchers in the related works did. Finally, it might be worth it to understand how to diversify the applicant pool as these models can only be as fair as the applicant pool allows them to be.

Overall, while the model has certain limitations and there exists potential steps to be taken for future improvement, the model outlined in this paper provides a promising basis for improving the White House's hiring process.

References

Prohibited employment policies/practices. n.d. US EEOC.

Agarwal Alekh, Beygelzimer Alina, Dudík Miroslav, Langford John, Wallach Hanna. A Reductions Approach to Fair Classification. 2018.

Alvero A., Arthurs N., Antonio A. L., Domingue B. W., Gebre-Medhin B., Giebel S., Stevens M. L. AI and Holistic Review: Informing Human Reading in College // ACM Digital Library. February 2020.

Kim M. P., Korolova A., Rothblum G., Yona G. Preference-Informed Fairness // arXiv. September 12 2019.

The United States Government . + algorithmic discrimination protections. October 4 2022a.

The United States Government . Presidential Personnel Office. September 16 2022b.

Veldanda A. K., Grob F., Thakur S., Pearce H., Tan B., Karri R., Garg S. Investigating Hiring Bias in Large Language Models // Open Review. n.d.

Zehlike Meike, Yang Ke, Stoyanovich Julia. Fairness in Ranking, Part I: Score-Based Ranking // ACM Comput. Surv. dec 2022. 55, 6.