Bias in Machine Learning Algorithms: A Critical Analysis of AI Systems

Jacob Derbyshire
University of New England
Science Faculty – SCI310
jderbysh@myune.edu.au

Disclaimer on the use of Artificial Intelligence

In the creation of this essay, the author utilised artificial intelligence (AI) language models as a supportive tool. The tool was used to aid in exploring initial concepts, analyse literature and to refine sentence construction and structure. The final content, including all arguments, analysis, and conclusions, represent the author's own work.

1. INTRODUCTION

This essay explores the use of artificial intelligence hiring systems (AHSs) (Sheard, 2025) and their use in hiring processes in Australian companies and across the world.

AHS tools are gaining in use and popularity as businesses seek to reduce costs and maximise efficiency, especially in cost centres such as human resources and recruitment functions. A 2019 industry survey reported that 88% of organisations globally have experimented with AI in recruitment activities (Mujtaba & Mahapatra, 2024).

This analysis is focused on three key areas:

First, algorithm facilitated discrimination introduced by model bias when used in resume screening by an AHS and disadvantageous effect this has on group members due to personal attributes such as gender, ethnicity and disability (Sheard, 2025). Legal considerations for businesses, which risk breaching regulations that protect these characteristics during the hiring process.

Secondly, an experiment conducted using ChatGPT 5.1 to validate earlier findings of bias and confirm whether this bias still applies to later models released when using a simulated AHS to evaluate resumes.

Finally, the feedback loop created by bias in AHSs and the continuing impacts that this may have on certain groups of people and businesses.

2. RESUME SCREENING

A major risk of using AHS tools for resume screening is model bias, where the system shows preference to certain groups based on characteristics unrelated to job performance. This is particularly significant in Australia, where characteristics such as race, sex, age, and religion are legally protected from discrimination (Ombudsman, 2025).

Following experiments conducted using large language model (LLM) AHS tools based on ChatGPT Wilson & Caliskan conclude that White and Male names are preferred, in 85.1% of 27 tests, the tool showed a preference for White names over Black names. Male names were preferred to Female names in 88.9% of the tests (Wilson & Caliskan, 2024).

From a commercial perspective, ignoring such biases has tangible costs. Businesses that deploy biased hiring models risk prosecution by regulators such as the Australian Human Rights Commission if their decision-making is proven to be discriminatory (Commission, 2025). Moreover, the resulting public scrutiny can lead to significant reputational damage and financial loss.

There are already several documented examples of this. In 2015, Google's job recommendation system was found to be showing gender bias by recommending hi-income job postings more frequently to men than to women. In 2019, Facebook's job ad system was shown to be skewed based on users' race and gender (Mujtaba & Mahapatra, 2024).

Further to this point, AHSs can encode proxies for protected characteristics that can be used to unfairly discriminate against certain groups of people. For example, if a model assigns a weighting to employment gaps on a resume, which could be resultant from maternity leave and therefore serve as a proxy for sex or gender (Sheard, 2025).

3. EXPERIMENT

To empirically test whether the biases identified by Wilson & Caliskan (2024) persist in newer models, an experiment was designed and conducted using the recently released ChatGPT 5.1.

The objective was to measure whether the model exhibited bias based on the racial and gender characteristics inferred from names on resumes. The experimental design drew upon the classic labour market discrimination study by Bertrand and Mullainathan (2003), utilising their list of 36 names statistically associated with African-American and White applicants.

8,748 samples were generated for analysis. To ensure real-world applicability the experiment focused on three entry-level industries: hospitality, administration and call centre work. For each industry, three different job advertisements were sourced from seek.com.au and anonymised, three corresponding resumes were then generated using AI to be a strong match for the roles (these are referred to as the target resumes). In total, there were nine job advertisements and nine resumes across the three industries. Each unique combination of job ad, resume, and one of the 36 names was repeated three times.

Utilising the OpenAI API, the model was then prompted to score each resume, out of 25, against four criteria: relevant experience, skills, achievements, and resume quality, providing a total score out of 100.

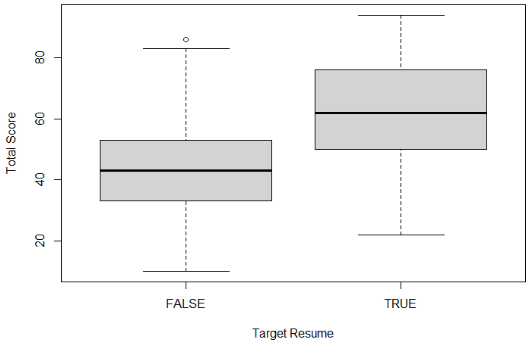

Figure 1: side-by-side boxplots showing score by target or non-target resume

To validate the integrity of the scoring system, the model's ability to differentiate between target and non-target resumes was analysed.

Figure 1 shows target resumes had a mean score of 61.54/100 compared to non-target resumes which had a mean score of 43.38/100. This is a statistically significant difference with a 95% confidence interval showing a difference in score of between 17.45 and 18.86, indicating that the AI scoring is meaningful and not arbitrary.

The primary analysis however, found no evidence of bias based on protected characteristics.

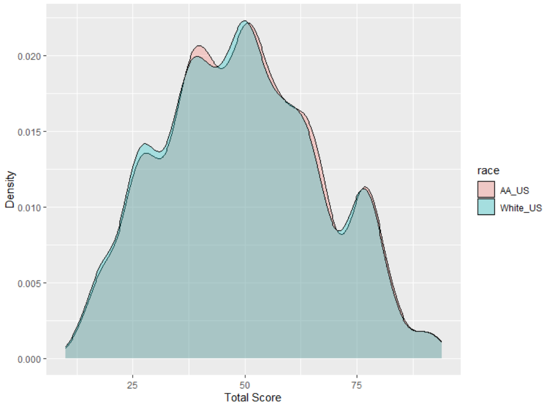

Figure 2: density plot showing density of total score by codified race

The density plot in figure 2 shows no significant difference between the names codified as being representative of African-American and White heritage and the evaluation score provided by the model.

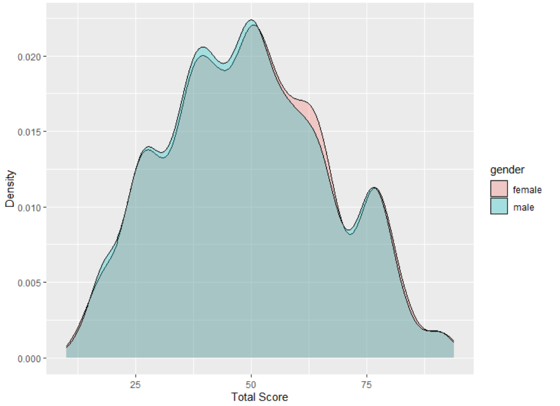

Figure 3: density plot showing density of total score by codified gender

Figure 3 shows there was no significant difference between the genders that could be inferred from the names and the evaluation score provided by the model.

While the scope of the study was limited, the results are notable. Within the tested parameters, ChatGPT 5.1 did not exhibit the same racial or gender biases in resume screening that were identified in earlier studies (Wilson & Caliskan, 2024). This suggests potential improvement in later models.

4. FEEDBACK LOOP

Sheard argues that data bias is perhaps the most common source of algorithm-facilitated discrimination, as data are not neutral. Sheard proposes that when training data are not truly representative of population data the models can start to embed real world discrimination (Sheard, 2025).

An example of this is in 2017, Amazon's AI candidate evaluation tool was found to be attributing lower scores to women's resumes as women were underrepresented in the training data (Mujtaba & Mahapatra, 2024).

A deployed AHS can create feedback loops where the model's decisions begin to influence its future training data. This is further reinforced when decision-makers treat the AI as an infallible substitute for human oversight.

These feedback loops can lead to a lack of diversity, sometimes colloquially referred to as 'Pale, Male and Stale'. The U.S. Supreme Court, in *Grutter v. Bollinger* (2003), argued that the competencies required for the modern global economy can only be cultivated through meaningful engagement with a wide spectrum of people, cultures, and viewpoints (Grutter v. Bollinger, 539 U.S. 306, 2003). As such, there is a clear negative impact to businesses as a result of reinforcement of biases.

For the individual, however, feedback loops can create systemic barriers that can unfairly limit their career opportunities and economic mobility.

5. SUMMARY

In conclusion, while Automated Hiring Systems (AHSs) are now prevalent in modern recruitment (Mujtaba & Mahapatra, 2024), their foundational models carry a significant risk of bias towards protected characteristics, as evidenced by high-profile cases at Amazon and Google.

This essay's experiment explored whether a contemporary AI LLM exhibits similar biases, finding no evidence of discrimination based on racial or gender characteristics inferred from candidate names. However, this specific finding does not suggest the model is free from bias. As data is never neutral (Sheard, 2025), it merely demonstrates that bias was not present in this particular paradigm.

The risk remains that flawed AHS recommendations can create compounding feedback loops if not properly scrutinised. This underscores a critical takeaway: AI is a tool to assist, not replace, human oversight, especially in decisions with a material impact on an individual's career and opportunities.

REFERENCES

Bertrand, M., & Mullainathan, S. (2003). *Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*. Cambridge, MA: National Bureau of Economic Research. Retrieved from https://doi.org/10.3386/w9873

Commission, A. H. (2025). *Right to a discrimination-free workplace*. Retrieved from Australian Human Rights Commission: https://humanrights.gov.au/our-work/right-discrimination-free-workplace

Derbyshire, J. (2025). *AIHiringBias (Version 1.0)*. Retrieved from GitHub: https://github.com/jderbysh/AIHiringBias

Grutter v. Bollinger, 539 U.S. 306 (U.S. Supreme Court 2003). Retrieved from https://supreme.justia.com/cases/federal/us/539/306/

Mujtaba, D. F., & Mahapatra, N. R. (2024). *Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions.* arXiv. Retrieved from https://arxiv.org/abs/2405.19699

Ombudsman, F. W. (2025). *Protection from discrimination at work*. Retrieved from Fair Work: https://www.fairwork.gov.au/employment-conditions/protections-at-work/protection-from-discrimination-at-work

Sheard, N. (2025). Algorithm-facilitated discrimination: A socio-legal study of the use by employers of artificial intelligence hiring systems. *Journal of Law & Society*. Retrieved from https://doi.org/10.1111/jols.12535

Wilson, K., & Caliskan, A. (2024). Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Retrieved from https://doi.org/10.1609/aies.v7i1.31748