Bias in AI-based recruitment tools

Frida Hartman[a,*]

[a]Faculty of Science, University of Helsinki, Finland
ORCID (Frida Hartman): https://orcid.org/0009-0002-3949-4952

Abstract. The aim of this research is to explore, detect and eventually mitigate bias that may arise as a result of the current use of artificial intelligence (AI) based tools in recruitment processes in the Nordics. By mapping the use of AI-based tools in recruitment, and the fairness or unfairness they display, recruitment can become both less biased and more efficient and accurate. It is important that the use of AI-based systems for decision-making is constantly questioned and monitored, to ensure ethical development of AI.

In this project, current bias detection and measurement tools are analysed to understand their advantages and possible disadvantages. The focus is on how well they detect and measure real-life bias that affect different stakeholders. The use of AI-based recruitment tools in the Nordics is investigated, both in terms of type, amount, and justification. Moreover, selected AI-based recruitment tools are tested using automatically generated job candidates to detect any bias that might exist in these tools. By gaining more insights into the situation in the Nordics, fairness in AI models can be promoted in a wider context.

1 Background

AI based systems are increasingly used in different fields. As the use of AI increases, so does the need for ethical and fair AI development. By establishing robust ethical and legal frameworks for AI tools [9], humans may be able to use them to make more informed and unbiased decisions.

Previous studies have revealed unfairness within many AI-based systems. Several AI-based facial recognition software have been shown to recognise white people better than black people [18] (racial bias) and men better than women [2] (gender bias). When considering intersectionality, i.e., how multiple personal attributes contribute to differential treatment, the performance difference becomes even greater when comparing, e.g., white men (error rate 0.0% - 0.3%) with black women (error rate 20.8% - 34.7%) [5]. Another example is the AI-based CV-scanning tool used by Amazon, which systematically favoured men over women [6]. These examples are results of structural and systematic bias that favour some groups over others.

The EU AI act (Regulation (EU) 2024/1689) categorises AI models into three risk categories; unacceptable risk (should not be used), high risk (should be regulated), and no/low risk (no need for regulation). Most of the models used for decision making are placed in the high-risk category. Among other areas, AI-based recruitment tools fall into the high-risk category, which means that they need to be regulated to determine whether the outcomes are fair or not. Given previous research that has shown bias in both facial recognition [5] and natural language processing (NLP) models [6], it is imperative to detect and mitigate bias in, e.g., CV-scanning tools and video-based sentiment and personality analysis to ensure fair treatment of job applicants.

There are claims that AI-based recruitment might help mitigate unfairness in manual recruitment [11]. However, many AI-based recruitment systems have been proven to display systematic bias against certain groups [4, 6]. Although AI systems might be less biased in some cases, intersectional subgroups are easily overlooked, leading to already vulnerable groups being discarded early in employment decisions [21]. Research has also shown that part of the problem with discrimination in AI systems is the lack of diversity among system developers [8, 23]. Thus, AI systems might help mitigate implicit or unconscious bias (biases that arise from internal subjective perceptions [3]) in manual recruitment. To do so, we must consider multifaceted types of systematic bias and discrimination, from the diversity issue among AI developers [8] to developing ethical frameworks for AI users [17].

In this PhD research, the aim is to use insights from both computer science and social sciences to understand bias from a multidisciplinary perspective. This will help computer scientists and other stakeholders make more inclusive and diverse decisions both while developing recruitment systems, but also while using them in real-world scenarios.

2 Research questions and scientific impact

The main objectives of this research are to minimise the bias in AI-based recruitment systems mainly in the Nordics, but with applicability elsewhere as well. The research questions to be addressed are

1. What are the most used and scientifically approved measurement and detection methods of bias in AI systems, and do these methods consider intersectional bias?
   - How well do existing methods detect bias, and how can they be improved?
2. What types of automated tools are used in recruitment in the Nordics?
   - How does the trade-off between accuracy and fairness manifest itself in the recruitment process, and is one favored over the other?
3. How well do AI-based recruitment tools work for different intersectional groups of people?
   - Are there specific groups that are often excluded or discriminated against, and in what ways?

* Corresponding Author. Email: frida.hartman@helsinki.fi

Detecting bias and understanding the underlying causes for bias in recruitment scenarios can lead to increased diversity in organisations, making work places more innovative and productive [15, 22]. Furthermore, fair AI models and techniques contribute to social and societal development that fosters equality and ethical decisions.

3 Research methods

This PhD project is based on three main articles, each aiming to address one of the research questions, as well as provide insight into the project as a whole.

The focus of the first article is on studying and comparing different bias detection and measurement methods. Since bias exists in most AI systems trained on data [14], it is essential to be able to measure bias, determine the type of bias, and understand the underlying mechanics of this bias. A pre-study for this first article is currently under review. In the pre-study, articles on bias detection in AI underwent an initial review according to inclusion and exclusion criteria. Some insights were uncovered on who conducts research on bias detection in AI and where this research is being done.

Based on observations from the pre-study, the scope was narrowed down to articles that somehow question purely technical bias detection, i.e. fairness metrics, or offer alternative approaches to detecting bias beyond fairness metrics. This decision to narrow down the search was also made to find bias detection methods that work for intersectional groups, since traditional fairness metrics might struggle to detect more complex forms of bias [7, 10].

The scope of the second article is to map the use of AI-based tools and automated systems in recruitment in the Nordics. To accomplish this, recruitment companies will be surveyed about their recruitment processes and their interest to participate further in the study. Based on how many companies volunteer to participate, either individual or group-based interviews are conducted. The main information needed from the participants is i) whether they use AI or not in their recruitment process, ii) if yes, what type of AI techniques are being used, and in what stage of the recruitment process, iii) if yes, how long have these methods been used at the organisation, iiii) what is the main reason for using the method that they are using, and iiiii) potential positive and negative experience from using the systems. Based on the answers and how many people they usually assist in employing, some conclusions can be drawn regarding the scope of using AI in recruitment processes in the Nordics.

In the third article, a scientific experiment will be conducted on AI-based recruitment tools selected from one or many of the companies surveyed for the second article. In the experiment, artificial job applications are generated using large language models (LLMs) and run through the selected tools. The aim of the experiment is to determine whether candidates belonging to historically discriminated groups and intersectional subgroups suffer from disadvantages in the recruitment process. Using artificial job applications, the system can be tested on a larger level, and protected attributes of the "applicant" are readily available without violating personal privacy. It should be noted that although artificially generated candidates provide many benefits, the reliability and accuracy of these candidates must be considered [1, 12].

4 State of the research

The pre-study is under review, and the results showed that proportionally more women are studying bias detection in AI compared to women studying other computer science subjects. It also showed that most of the research is being done in the US, but according to recent paradigm shifts [13, 20], it remains to be seen if this will be the case in some years.

Based on observations from the pre-study, namely that there is somewhat of an automation bias (an excessive trust in technological solutions) [19], the remaining SLR will focus on the issues that arise when using a techno solutionist perspective [16] to solve a socio-technical problem. The SLR is currently in the writing phase and is expected to be submitted for review this year. After the SLR has been submitted, the focus shifts towards the second article.

5 Ethical considerations

In this project, empirical data will be collected through surveys and interviews. Participation in the study is voluntary, and all participants will be asked for explicit consent to take part in the the survey and possible follow-up interviews. Interviews are recorded, transcribed and analysed using qualitative and, if necessary, computational methods. Any collected data will be pseudonymised and should any respondents share information that could be considered secret or sensitive with regards to their company, these data will be removed from all documentation immediately. Participants will be informed about their right to withdraw from the study at any time.

Participants will be asked whether they want to share their responses with the scientific community outside of this project (participation is possible without data-sharing). If the amount of data collected that is allowed to be shared is deemed useful outside the context of this project, it will be shared with other researchers. All data will be stored according to the GDPR guidelines and handled according to Finnish guidelines for ethical research. Any open source data that is used will be accredited to the people or institution that collected it.

6 Timetable

This PhD project is planned to span four years (1.8.2024-30.7.2028). During the first year, the SLR was conducted and is planned to be submitted for publication in the beginning of the second academic year. During the second year, the survey for the second article is conducted. Potential organisations that fit the criteria for the survey are mapped and contacted during the autumn term. During the spring term, survey results are interpreted and further interviews are conducted. At the start of the third academic year, the second article is written and submitted for publication. Also, the tests to detect bias in one (or a few) recruitment tools are conducted and analysed during the rest of the third year. At the beginning of the fourth year, the results are summarised and the third article is submitted for publication. During the last term, the PhD project is summarised, and the dissertation is defended. Table 1 presents a brief timetable of the planned work and publications in this project.

Table 1. Timetable and publication plan. Past events are marked in gray.

Time period	activity	publication
A 2024	Work on SLR	-
S 2025	Work on SLR	submitted pre-study
A 2025	Work on survey	submit SLR
S 2026	Work on survey	-
A 2026	Work on survey	submit survey
S 2027	Work on recruitment test	-
A 2027	Work on recruitment test	submit recruitment test
S 2028	Summarise PhD	defend dissertation

References

[1] A. Amirova, T. Fteropoulli, N. Ahmed, M. R. Cowie, and J. Z. Leibo. Framework-based qualitative analysis of free responses of large language models: Algorithmic fidelity. *Plos one*, 19(3):e0300024, 2024.

[2] M. Atay, H. Gipson, T. Gwyn, and K. Roy. Evaluation of gender bias in facial recognition with traditional machine learning algorithms. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE, 2021.

[3] G. Beattie and P. Johnson. Possible unconscious bias in recruitment and promotion and the need to promote equality. *Perspectives: Policy and Practice in Higher Education*, 16(1):7–13, 2012.

[4] M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn, December*, 7, 2018.

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[6] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

[7] A. DeVos, A. Dhabalia, H. Shen, K. Holstein, and M. Eslami. Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19, 2022.

[8] E. Drage and K. Mackereth. Does ai debias recruitment? race, gender, and ai's "eradication of difference". *Philosophy & technology*, 35(4): 89, 2022.

[9] E. Ferrara. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, 6(1):3, 2023.

[10] A. Ghosh, L. Genuit, and M. Reagan. Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR, 2021.

[11] K. A. Houser. Can ai solve the diversity problem in the tech industry: Mitigating noise and bias in employment decision-making. *Stan. Tech. L. Rev.*, 22:290, 2019.

[12] S. Lee, T.-Q. Peng, M. H. Goldberg, S. A. Rosenthal, J. E. Kotcher, E. W. Maibach, and A. Leiserowitz. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8):e0000429, 2024.

[13] J. Mervis. Trump orders cause chaos at science agencies. *Science (New York, NY)*, 387(6734):564–565, 2025.

[14] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163, 2021.

[15] M. Mokhtech, R. Jagsi, R. M. Vega, D. W. Brown, D. W. Golden, T. Juang, M. D. Mattes, C. C. Pinnix, and S. B. Evans. Mitigating bias in recruitment: attracting a diverse, dynamic workforce to sustain the future of radiation oncology. *Advances in Radiation Oncology*, 7(6): 100977, 2022.

[16] L. Naudts, A. Vedder, and N. Smuha. Fairness and artificial intelligence. 2025.

[17] O. L. Olorunfemi, O. O. Amoo, A. Atadoga, O. A. Fayayola, T. O. Abrahams, and P. O. Shoetan. Towards a conceptual framework for ethical ai development in it systems. *Computer Science & IT Research Journal*, 5(3), 2024.

[18] B. Orr, A. Sumsion, S. Torrie, and D.-J. Lee. Exploring racial bias in deep face recognition models. In *2024 Intermountain Engineering, Technology and Computing (IETC)*, pages 92–97, 2024. doi: 10.1109/IETC61393.2024.10564236.

[19] S. Strauß. Deep automation bias: how to tackle a wicked problem of ai? *Big Data and Cognitive Computing*, 5(2):18, 2021.

[20] The White House. Ending radical and wasteful government dei programs and preferencing, 2025. URL https://www.whitehouse.gov/presidential-actions/2025/01/ending-radical-and-wasteful-government-dei-programs-and-preferencing/. (accessed: 06.05.2025).

[21] N. Tilmes. Disability, fairness, and algorithmic bias in ai recruitment. *Ethics and Information Technology*, 24(2):21, 2022.

[22] R. Vivek. Enhancing diversity and reducing bias in recruitment through ai: a review of strategies and challenges. *Informatics. Economics. Management*, 2(4):0101–0118, 2023.

[23] S. M. West, M. Whittaker, and K. Crawford. Discriminating systems. *AI Now*, 2019:1–33, 2019.