Mitigating diversity biases of AI in the labor market

Rigotti, C.; Puttick, A.; Fosch Villaronga, E.; Kurpicz-Briki, M.; Alvarez, J.M.; Fabris, A.; ... ; Zehlike, M.

Mitigating Diversity Biases of AI in the Labor Market

Carlotta Rigotti[1], Alexandre Puttick[2], Eduard Fosch-Villaronga[1] and Mascha Kurpicz-Briki[2]

[1]Leiden University, Rapenburg 70, 2311 EZ Leiden, Netherlands
[2]Berner Fachhochschule BFH, Technik und Informatik, Quellgasse 21, 2501 Biel, Switzerland

Abstract

In recent years, artificial intelligence (AI) systems have been increasingly utilized in the labor market, with many employers relying on them in the context of human resources (HR) management. However, this increasing use has been found to have potential implications for perpetuating bias and discrimination. The BIAS project kicked off in November 2022 and is expected to develop an innovative technology (hereinafter: the Debiaser) to identify and mitigate biases in the recruitment process. For this purpose, an essential step is to gain a nuanced understanding of what constitutes AI bias and fairness in the labor market, based on cross-disciplinary and participatory approaches. What follows is a preliminary overview of the design and expected implementation of the project, as well as how our project aims to contribute to the existing literature on law, AI, bias, and fairness.

Keywords

bias, fairness, discrimination, trustworthy AI, labor market, natural language processing

1. Setting the scene: The increasing deployment of AI application in the labor market

Employers are integrating AI applications into their HR processes to increase efficiency and cut costs [1, 2], with the lockdown measures in response to the Covid-19 pandemic accelerating this process for public health protection [3]. AI applications may be deployed in the labor market in order to identify the best job applicant or monitor and manage the behavior and performance of workers; they can design job descriptions, market vacancies to potential candidates, provide technical support through chatbots, analyze large volumes of applicant data, and conduct interviews [4]. Furthermore, they can track, train, and rate workers, monitor infractions leading to disciplinary proceedings, and predict resignation or health problems [5, 6, 7, 8].

Against this backdrop, Ajunwa [9] defines the deployment of AI applications in the labor market as 'a black box at work,' which lacks transparency, accountability or explanation about monitoring practices. More generally, a growing body of literature voices concern about privacy [10, 3, 8] and fairness [11], in the sense that AI applications are likely to perpetuate social marginalization and discrimination. Prejudices and flaws in past datasets and human programmers can indeed easily seep into code [12, 13].

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
c.rigotti@law.leidenuniv.nl (C. Rigotti); alexandre.puttick@bfh.ch (A. Puttick); e.fosch.villaronga@law.leidenuniv.nl (E. Fosch-Villaronga); mascha.kurpicz@bfh.ch (M. Kurpicz-Briki)
0000-0001-8956-0677 (C. Rigotti); 0000-0002-2142-0309 (A. Puttick); 0000-0002-8325-5871 (E. Fosch-Villaronga); 0000-0001-5539-6370 (M. Kurpicz-Briki)

2. Gaining new knowledge about AI bias and fairness in the labor market

Addressing a lack of consensus about the interpretation of AI bias and fairness in the labor market, we begin our research with a systematic literature review. In brief, we draw on the traditional distinction of individual fairness *versus* group fairness [14], disparate treatment *versus* disparate impact [15, 16, 12], pre-existing, technical, and emergent bias [17] *versus* other shapes and forms, including representation bias arising from the way we define and sample a population and temporal bias stemming from differences in populations and behaviors over time [18]. In doing so, we acknowledge that the nature and significance of bias and fairness in humans versus in AI do not overlap, to the extent that some research radically suggests the unsuitability of mathematics to capture the full meanings of these social concepts [19]. For Abu-Elyounes [14], AI fairness is contextual and requires balancing between competing values on a case-by-case basis, necessitating specialized consideration of the employment sector within the BIAS project.

In validating and contextualizing the research outputs of the literature review, we also perform some fieldwork. To be precise, we have interviewed 70 HR managers and AI developers, who are located in Estonia, Iceland, Italy, the Netherlands, Norway, Switzerland, and Turkey, to ascertain their views on the use of AI applications in the labor market and its impact on fairness. Briefly, it appears that most respondents have a positive attitude towards the deployment of AI applications for recruitment and selection purposes, while voicing concern about the management of staff. Although lacking a common understanding of fairness, a number of requirements are listed for AI-driven human resources practices to be considered fair, especially in terms of human oversight and transparency. Also, respondents call for the adoption of mitigation measures to address diversity biases, with special emphasis on gender bias. Besides AI-driven solutions similar to the Debiaser, they cite the examples of diversity quotas, DI officers, specific guidelines, and training on the matter. At the same time, we have drafted a Qualtrics survey of workers' attitudes towards AI, the aim of which is to understand workers' experiences and opinions of AI applications used in an HR management context and how they might lead to bias and discrimination. Although the survey has just been launched and will remain open until July, it has been reported that the majority of respondents have a fairly negative attitude towards the use of AI applications in the labor market. In any case, for AI-driven HR practices to be considered fair, they believe in the provision of equal opportunities despite the personal characteristics of the individual as well as the transparency and explicability of the technology making a specific decision.

3. Designing the 'Debiaser'

Based on the literature review and the fieldwork, the next step of the BIAS project will be the design and piloting of the Debiaser. In doing so, we contribute to an existing body of research suggesting possible remedies for AI bias in both social and computer science [18]. The Debiaser is conceptualized as a tool for bias detection and mitigation in the context of a common use case for text-based AI recruitment tools—the automatic pre-selection of job applicants for first-round

interviews. The tool will be built upon the principles of trustworthy AI [20] and incorporate the rich knowledge base assembled by consortium partners in the framework of the BIAS project. A major contribution of our work will be the extension and adaptation of existing methods to EU regional languages and local prejudices, with a focus on specialized methods targeting fair recruitment practices.

The scientific research we conduct explores weak points in existing AI recruitment tools. The first concerns off-the-shelf word embeddings such as Word2Vec [21] and pre-trained large language models (PLMs) such as BERT [22] or GPT-3 [23]. Highly capable text processing models require extensive resources in terms of time, computing power and data, and thus many AI applications, including recruitment tools, are built using such word embeddings or PLMs. These machine-learning-based methods often reproduce prejudices existing in their training corpora, with a notorious example being Amazon's own (now scrapped) job candidate selection tool, which was demonstrated to systematically discriminate against women [24]. We aim to utilize and further develop methods for detecting, measuring and mitigating bias in word embeddings and PLMs, such as those introduced by Caliskan et al. [25], Guo and Caliskan [26] and Ahn and Oh [27]. This builds upon previous work which extends the methods of Caliskan et al. [25] to French, German, Italian and Swedish word embeddings [28, 29].

Most current AI applications make use of black-box neural network models whose decision-making process is very difficult to distill into human-understandable explanations. This inhibits trust in such systems and complicates the task of auditing algorithmic decisions. To this end, we aim to utilize state-of-the-art methods in explainable natural language processing (XNLP) [30], aiming to provide clear explanations for all stages of bias detection and mitigation deployed within the Debiaser. The XNLP methods we will integrate can be sorted into two basic approaches toward improving the interpretability of algorithmic decisions: Feature-selection methods [31, 32, 33] aim to identify which words or phrases in the input data played an important role in a given model's output decision. The second category of techniques generate natural language explanations which should explain the model's decision in comprehensible, plain-language terms [34, 35]. In addition, our research will target the detection of text that is susceptible to unfair discrimination and explore the potential of suggested rewrites as a technique for avoiding such discrimination, using reinforcement learning-based techniques in the same vein as Sharma et al. [36]. Both the suggested rewrites and natural language explanations generated by the Debiaser can be further improved using human feedback to further train the underlying models with the goal of generating output corresponding to user preferences [37, 38].

4. Conclusion

The growth of the role of AI in HR management is to continue, but its vast potential comes with various challenges. AI applications are far from being free from bias and discrimination, and AI developers and HR executives should be prepared to grapple with difficult questions and ensure that this technology is implemented in a responsible manner. For this purpose, the BIAS project aims at identifying and mitigating diversity biases in the labor market, by combining cross-disciplinary fieldwork and designing new technological solutions.

Acknowledgments

This work is part of the Europe Horizon project BIAS funded by the European Commission, and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

[1] N. Guenole, S. Feinzig, The business case for ai in hr: With insights and tips on getting started, IBM Smarter Workforce Institute (2018).

[2] N. T. Tippins, F. L. Oswald, S. M. McPhail, Scientific, legal, and ethical concerns about ai-based personnel selection tools: a call to action, Personnel Assessment and Decisions 7 (2021) 1.

[3] I. Ebert, I. Wildhaber, J. Adams-Prassl, Big data in the workplace: Privacy due diligence as a human rights-based approach to employee privacy protection, Big Data & Society 8 (2021) 20539517211013051.

[4] M. F. Gonzalez, W. Liu, L. Shirase, D. L. Tomczak, C. E. Lobbe, R. Justenhoven, N. R. Martin, Allying with ai? reactions toward human-based, ai/ml-based, and augmented hiring processes, Computers in Human Behavior 130 (2022) 107179.

[5] M. V. V. Yawalkar, a study of artificial intelligence and its role in human resource management, International Journal of Research and Analytical Reviews (IJRAR) 6 (2019) 20–24.

[6] S. M. C. Loureiro, J. Guerreiro, I. Tussyadiah, Artificial intelligence in business: State of the art and future research agenda, Journal of business research 129 (2021) 911–926.

[7] B. Eubanks, Artificial intelligence for HR: Use AI to support and develop a successful workforce, Kogan Page Publishers, 2022.

[8] A. Aloisi, Regulating algorithmic management at work in the european union: Data protection, non-discrimination and collective rights, International Journal of Comparative Labour Law and Industrial Relations, Forthcoming (2022).

[9] I. Ajunwa, The "black box" at work, Big Data & Society 7 (2020) 2053951720966181.

[10] D. P. Bhave, L. H. Teo, R. S. Dalal, Privacy at work: A review and a research agenda for a contested terrain, Journal of Management 46 (2020) 127–164.

[11] A. L. Hunkenschroer, C. Luetge, Ethics of ai-enabled recruiting and selection: A review and research agenda, Journal of Business Ethics 178 (2022) 977–1007.

[12] J. Kleinberg, J. Ludwig, S. Mullainathan, C. R. Sunstein, Discrimination in the age of algorithms, Journal of Legal Analysis 10 (2018) 113–174.

[13] M. Raub, Bots, bias and big data: artificial intelligence, algorithmic bias and disparate impact liability in hiring practices, Ark. L. Rev. 71 (2018) 529.

[14] D. Abu-Elyounes, Contextual fairness: A legal and policy analysis of algorithmic fairness, U. Ill. JL Tech. & Pol'y (2020) 1.

[15] S. Barocas, M. Hardt, A. Narayanan, Fairness in machine learning, Nips tutorial 1 (2017) 2017.

[16] M. MacCarthy, Standards of fairness for disparate impact assessment of big data algorithms, Cumb. L. Rev. 48 (2017) 67.
[17] B. Friedman, H. Nissenbaum, Bias in computer systems, ACM Transactions on Information Systems (TOIS) 14 (1996) 330–347.
[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Computing Surveys (CSUR) 54 (2021) 1–35.
[19] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 59–68.
[20] A. Hleg, Ethics guidelines for trustworthy ai, B-1049 Brussels (2019).
[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).
[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[23] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[24] J. Dastin, Amazon scraps secret ai recruiting tool that showed bias against women, in: Ethics of data and analytics, Auerbach Publications, 2018, pp. 296–299.
[25] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.
[26] W. Guo, A. Caliskan, Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 122–133.
[27] J. Ahn, A. Oh, Mitigating language-dependent ethnic bias in bert, arXiv preprint arXiv:2109.05704 (2021).
[28] M. Kurpicz-Briki, Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings (2020).
[29] M. Kurpicz-Briki, T. Leoni, A world full of stereotypes? further investigation on origin and gender bias in multi-lingual word embeddings, Frontiers in big Data 4 (2021) 625290.
[30] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable ai for natural language processing, arXiv preprint arXiv:2010.00711 (2020).
[31] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[32] S. Rathi, Generating counterfactual and contrastive explanations using shap, arXiv preprint arXiv:1906.09293 (2019).
[33] J. Li, X. Chen, E. Hovy, D. Jurafsky, Visualizing and understanding neural models in nlp, arXiv preprint arXiv:1506.01066 (2015).
[34] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, P. Blunsom, e-snli: Natural language inference with natural language explanations, Advances in Neural Information Processing Systems 31 (2018).
[35] S. Narang, C. Raffel, K. Lee, A. Roberts, N. Fiedel, K. Malkan, Wt5?! training text-to-text

models to explain their predictions, arXiv preprint arXiv:2004.14546 (2020).
[36] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, T. Althoff, Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach, in: Proceedings of the Web Conference 2021, 2021, pp. 194–205.
[37] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, Advances in neural information processing systems 30 (2017).
[38] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, G. Irving, Fine-tuning language models from human preferences, arXiv preprint arXiv:1909.08593 (2019).