Towards Relational Egalitarianism in the Third Wave of Algorithmic Justice

Nowadays, algorithmic decision-making (ADM) is used for various high-stake applications with fundamental impacts for the decision subjects. Consequential decisions about loans, crime sentencing, or hiring may potentially impact the life of every civil person. However, examples of unfairly biased decisions and discrimination – such as the gender-biased hiring algorithm of Amazon – raised concerns regarding algorithmic justice.

Within an interdisciplinary research area, several formal algorithmic fairness metrics have been developed to approach these injustices, mainly based on a distributive understanding of justice (Kasirzadeh, 2022). In the developed fairness-aware machine learning research three waves can be identified (Huang et al., 2022; Kind, 2020). The 1st wave is dominated by the discussion about appropriate guidelines and principles to ensure a fair development and use of ADM. To account for these guidelines, the 2nd of research made considerable efforts to develop a mathematical solution in order to identify and mitigate unfair biases. Yet, these technical fairness metrics received more and more critique because of several conceptual problems. The critique on the technical approaches evoked the 3rd wave of AI fairness, which emphasizes that algorithms have to be seen as socio-technical systems. While the 2nd wave is concerned with distributive questions of the decision outcome (e.g., What is the share between men and women who received a job interview invitation?), the 3rd wave has a broader focus, including power dynamics and social structures (Kind, 2020).

The development from the second to the third wave within the fair-ML research shows similarities to the philosophical discourse on justice, which can be subdivided into distributive accounts and relational accounts of justice. Distributive approaches, focusing on different currencies of equality (e.g., income, wealth, resources) and how they ought to be distributed, are opposed by relational accounts, which conceptualize equality based on the quality of social relations among citizens and the treatment of citizens by social institutions, focusing on unequal power asymmetries, social relations, and structural injustices (Arneson, 2013).

This presented work incorporates the relational view into the discussion of algorithmic justice by analyzing the implications of the 3rd wave of AI fairness for ADM in the application example of hiring. Therefore, the following research question is investigated: *Which topics, including critical and constructive approaches, are discussed within the literature on relational*

algorithmic justice and what are its implications for automated decision-making in hiring? Accordingly, the aim of this work is twofold: On the one hand, it synthesizes the contents of the 3rd wave of algorithmic justice and its relational implications, and on the other hand, it conducts a context-specific evaluation for the case study hiring. To this end, the presented thesis follows an interdisciplinary method based on a systematic literature review (SLR) supplemented by normative reasoning.

While several reviews regarding the first wave (Gaelle Cachat-Rosset & Alain Klarsfeld, 2023; Jobin et al., 2019) and the second wave (e.g., Caton & Haas, 2020; Pessach & Shmueli, 2022) exist, there are no systematic reviews of the emerging third wave yet. Häußermann and Lütge (2022) highlight that this third wave is currently evolving through critical contributions of the previous approaches to algorithmic justice and does not yet have a clear upshot. This research gap motivates the SLR, aiming to draw a clear picture on these latest advances in the research on algorithmic justice.

Hypothesizing that the third wave of algorithmic justice is highly influenced by the thematic discourses of relational justice, the SLR crystallized insights of 29 scientific contributions to develop the notion of 'relational algorithmic justice'. The search query was based on theories on relational egalitarianism (Anderson, 1999; Nath, 2020; Young, 1990) and enriched with technical keywords such as 'Machine Learning' or 'algorithm'. The results were distilled to compile underexplored topics within the distributive frame of algorithmic justice and analyses the underlying motivation, the primarily discussed topics, and critical and constructive approaches of the third wave. Particularly evident became the critical emphasis on the categorization and measurements of humans; the interplay between algorithms, power, and capitalism; and epistemic challenges of algorithmic fairness.

Striving for a context-based discussion, the identified implications from relational algorithmic justice are analysed for the moral evaluation of value-laden ADM in the case study hiring. The identified topics are used to structure the discussion of implications for the use of ADM in hiring. All in all, the developed notion of '*relational algorithmic justice'* highlights that the prevailing orientation of algorithmic fairness research towards strict egalitarian principles and questions of distributions is not suitable to account for structural injustices. Following a philosophical reasoning based on the commitment towards the unconditional appreciation of human diversity, it is revealed that ADM in hiring fails to recognize individual differences. When it comes to high-stakes decisions that involve value-laden considerations, i.e., decisions that are not based

on facts alone but must be morally justified, a relational approach to algorithmic fairness is crucial to ensure that the impacts of automated decision-making are equitable for all.

References

Anderson, E. (1999). What is the Point of Equality? *Ethics*, *109*(2), 287–337. http://www.jstor.org/stable/2989479

Arneson, R. (2013). Egalitarianism. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/egalitarianism/

Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey. *Cornell University*. https://arxiv.org/abs/2010.04053

Gaelle Cachat-Rosset, & Alain Klarsfeld (2023). Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines. *Applied Artificial Intelligence*, *37*(1), 2176618. https://doi.org/10.1080/08839514.2023.2176618

Häußermann, J. J., & Lütge, C. (2022). Community-in-the-loop: Towards pluralistic value creation in AI, or—Why AI needs business ethics. *AI and Ethics*, *2*(2), 341–362. https://doi.org/10.1007/s43681-021-00047-2

Huang, L. T.-L., Chen, H.-Y., Lin, Y.-T., Huang, T.-R., & Hun, T.-W. (2022). Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy. *Feminist Philosophy Quarterly*, *8*(3/4).

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kasirzadeh, A. (2022). Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *AIES '22: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery (ACM)* (pp. 348–356). https://doi.org/10.1145/3514094.3534188

Kind, C. (2020, August 23). The term 'ethical AI' is finally starting to mean something. *VentureBeat*. https://venturebeat.com/ai/the-term-ethical-ai-is-finally-starting-to-mean-something/

Nath, R. (2020). Relational egalitarianism. *Philosophy Compass*, *15*(7). https://doi.org/10.1111/phc3.12686

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3), 1–44.

Young, I. M. (1990). Displacing the Distributive Paradigm (Chapter 1). In I. M. Young (Ed.), *Justice and the politics of difference* (pp. 15–38). Princeton University Press. https://www.degruyter.com/document/doi/10.1515/9781400839902-004/html?lang=de