Beyond Distributions: A Systematic Review on Relational Algorithmic Justice

Laila Wegner[1,*] Marie Christin Decker[1,*] and Carmen Leicht-Scholten[1]

[1]RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany

Abstract

Examples of algorithmically reinforced inequalities motivated a growing research area on algorithmic fairness. Traditional fairness metrics mainly focus on distributive questions of fairness such as the distribution of positive outcomes within a protected group. However, in philosophy not only distributive but also relational accounts of justice exist, focusing on power hierarchies and structural inequalities. These topics are also the subject of the currently emerging third wave of algorithmic fairness, stressing that algorithms have to be seen as socio-technical systems. We aim to analyze the latest developments of the research in more detail and investigate what a relational perspective on justice adds to the (so far merely distributive) research on algorithmic fairness. Using a systematic literature review, we plan to focus on a novel perspective of relational algorithmic justice and highlight underexplored topics as well as critical and constructive approaches within the third wave of algorithmic fairness.

Keywords

Algorithmic Fairness, Relational Justice, Systematic Literature Review, Machine Learning, Algorithmic Decision-Making

1. Introduction

Examples of harmful algorithmic biases in high-stakes decisions led widely to efforts to reduce the potentially negative impacts of algorithmic decision-making (ADM). Within an interdisciplinary research area, several formal algorithmic fairness metrics have been developed, mainly based on a distributive understanding of justice [1]. However, in the philosophical discussion of justice, two families of justice stand out [2]: 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 [2], focusing on unequal power asymmetries, social relations, and structural injustices.

Aiming to design fair algorithmic decisions poses complex distributive questions, especially against the background of biased data and algorithms. The comparison of statistics such as error, true positives, or false positive predictions between different members of so-called protected attributes (e.g. gender, race, ...) has led to extensive research on different fairness metrics [e.g., 11] and challenges such as the 'impossibility theorem'. While it is as urgent as difficult to approach the distributive challenges of algorithmic biases, it may not directly lead to a holistic perspective of algorithmic justice. Within a merely distributive framework questions such as 'How does an ADM affect the interaction between the decision subject and decision maker?', 'Who benefits and who is harmed by the use of ADM in a specific context?', 'How does the power between decision subjects and an institution shift once an ADM is involved?', and 'How are those who use ADM (e.g., Recruiter)

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany

* Corresponding author.

laila.wegner@rwth-aachen.de (L. Wegner); marie.decker@gdi.rwth-aachen.de (M. Decker); carmen.leicht@gdi.rwth-aachen.de (C. Leicht-Scholten)

0000-0001-9983-8361 (L. Wegner); 0000-0002-9138-231X (M. Decker); 0000-0003-2451-6629 (C. Leicht-Scholten)

affected in their daily work?' remain almost unnoticed. This kind of question substantiates the demand to expand algorithmic fairness research with a relational focus.

Although the distributive perspective predominates the discourse on algorithmic justice, the currently emerging third wave [3, 4] puts attention to the drawbacks of this approach. In general, scholars of the third wave of AI ethics emphasize that algorithms have to be seen as socio-technical systems that necessitate the discussion of power dynamics and social structures when talking about algorithmic justice [4, 5]. The three waves of AI ethics were named for the first time by Carley Kind [5], director of the Ada Lovelace Institute, in an online blog post that is increasingly taken up by scientific literature [e.g., 4, 6–8]. Roughly, it can be summarized that the first wave was dominated by guidelines and principles that demanded fair development and use of algorithms (for reviews see [9, 10]). The second wave aimed to overcome the abstract and high-level character of guidelines and developed mathematical solutions to identify and mitigate unfair biases [e.g., 11, 12]. Häußermann and Lütge [3] highlight that the third wave is currently evolving and does not yet have a clear upshot. Thus, there are no systematic reviews of the emerging third wave yet. This research gap motivates our contribution, aiming to draw a clear picture of these latest advances in the research on algorithmic justice. Importantly, the mathematical fairness approaches developed in the second wave are mainly concerned with distributive questions of the algorithmic outcome such as the share between men and women getting a positive prediction. In contrast, the broader focus of the third wave includes power dynamics and structural inequalities and thus seems to focus on thematic discourses typical for relational theories of justice.

Agreeing with Branford's statement at the CEPE [13] that AI Ethics currently faces a "relational turn" which should be further encouraged, we followed our hypotheses that the third wave of algorithmic justice is highly influenced by the thematic discourses of relational justice. We aim to support the perspective of 'relational algorithmic justice'. To do so, we executed a systematic literature research (SLR) and crystallized insights of scientific contributions to the developing third wave of algorithmic justice.

2. Background: Structural Inequalities and Relational Justice

As highlighted by Binns [14], the developed algorithmic fairness metrics often lack an in-depth consideration of the moral foundations of justice. Especially relational justice is a rather unpopular perspective within the algorithmic fairness literature. Distributive and relational accounts share the basic agreement that every person has an equal moral worth. Thus, for both theories equality is a central aspect, often referred to as egalitarianism [2]. From then on, the theories differ on what should be the central concern of egalitarians – either focusing on how some goods such as resources, income, or wealth ought to be distributed (i.e., distributive justice) or focusing on social relationships, treatment with mutual respect, and power imbalances (i.e., relational justice). Relational egalitarians consider distributive injustices as a symptom caused by social injustices and thus demand to focus on the root cause such as social relations and structures instead. This means, relational justice might have also distributive implications but stresses that only focusing on distributive injustices may not display the whole picture.

Centering the root of inequalities means focusing on oppression, domination, and unequal power hierarchies. Equal relations demand treatment with reciprocity and mutual respect, with no one who perceives themselves as superior or inferior to others. This does not only refer to the interpersonal level but also to the structural level, considering how social institutions treat citizens. The structural level is fundamentally influenced by Iris Young [15] who coined the concept of structural inequalities. These inequalities result from a sum of non-blameworthy processes that limit the capabilities of large groups while others benefit by gaining power and privileges. The non-blameworthy processes refer to the rational decision to prioritize personal goals such as employees outsourcing labor to low-wage countries [16]. While it is a decision within the societal rules and norms and mainly motivated by

economic incentives, it is also an example of reinforced exploitive structures in which individuals contribute to structural injustices without bad intentions.

To summarize, the individual level of relational justice focuses on the treatment with mutual respect, and the institutional level highlights the structural nature of injustices. Both families of relational justice will inform the notion of relational algorithmic justice which is investigated in a systematic literature review.

3. Preliminary Results and Discussion

To the best of our knowledge, no systematic reviews corresponding to the third wave of algorithmic justice have yet been conducted. Based on the observation that the third wave of algorithmic justice centers relational topics, our systematic literature review follows a search query based on theories on relational egalitarianism [e.g., 15, 17, 18] enriched with technical keywords such as 'machine learning' or 'algorithm'. Following predefined inclusion and exclusion criteria, we will select papers focusing on relational concerns within the algorithmic justice research. Afterward, the included papers are analyzed in detail for critical and constructive approaches of the third wave.

First results indicate several underexplored topics within the distributive frame of algorithmic justice. Among others, particularly evident became the critical emphasis on the categorization and measurements of humans, stressing that the selection of protected attributes is subjective [e.g., 19] and reduces fluid concepts such as gender identity to discrete categories [20]. This oversimplification can lead to misrepresentation and stigmatization [21, 22]. Furthermore, the analyzed literature highlights the interplay between algorithms, power, and capitalism, including a critical analysis of the approaches to intersectionality. Here, it stands out that the current approaches to intersectional fairness focus on subgroup fairness while failing to engage with systems of oppression [19, 23]. Additionally, the relational focus revealed several epistemic challenges of algorithmic fairness, highlighting for example that the discourse of algorithmic fairness is Western centralized [24–26] and hard codes societal norms by treating constructed categories as facts [22, 27–29].

The primary findings of the literature review critically highlight several issues of current algorithmic fairness approaches. However, several authors also express the hope that algorithmic systems could be used proactively to reduce structural injustices. In this spirit, Kasirzadeh [1] states that "there is the potential that algorithmic systems can be used to repair some problematic structures and to generate better ones" (p. 355). Leavy et al. [30] suggest that this might be reached when quantitative data is actively used to highlight and combat racism. While emphasizing the difficult challenge of using algorithms as a driver for more justice, Zajko [31] summarizes that "there may be ways of designing new AI systems that help to shift power, as long as this is done with the participation of the people, groups, and communities that such efforts are intended to help" (p. 1048).

A challenge of the outlined SLR involves ensuring that the relational perspective on algorithmic fairness also contains concrete guidance for its implementation. However, this challenge lies in the nature of the presented subject; several authors criticize that research on challenges without a clear solution strategy is likely to be considered outside the scope of the research on algorithmic fairness [24,27,31]. Relational algorithmic justice must extend the focus beyond technical and clear solutions to consider complex structural dependencies. Thus, the strength of the presented SLR lies in the compilation of underexplored problems and critical reflections within algorithmic justice research – a necessary first step before founded guidance of relational algorithmic justice can be developed.

Summarizing, this literature review and a focus on relational algorithmic justice underlines the need to consider a broader scope than the identification and mitigation of biases within the algorithmic justice research. The relational perspective extends the focus beyond technical solutions to consider complex structural dependencies. An interdisciplinary, well-founded examination of philosophical approaches is essential to improve the efforts of algorithmic justice research and has

the potential to enable the use of algorithms for fighting structural injustices and enabling structural change.

References

[1] A. Kasirzadeh, 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), 2022, pp. 348–356. Accessed: Jan. 26 2023.

[2] R. Arneson, Egalitarianism, in: The Stanford Encyclopedia of Philosophy, Edward N. Zalta, Ed., 2013rd ed.: Metaphysics Research Lab, Stanford University, 2013. URL: https://plato.stanford.edu/entries/egalitarianism/

[3] J. J. Häußermann and C. Lütge, Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics, AI Ethics, vol. 2, no. 2, pp. 341–362, 2022, doi: 10.1007/s43681-021-00047-2.

[4] L. T.-L. Huang, H.-Y. Chen, Y.-T. Lin, T.-R. Huang, and T.-W. Hun, Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy, Feminist Philosophy Quarterly, vol. 8, ¾, 2022.

[5] C. Kind, The term 'ethical AI' is finally starting to mean something, VentureBeat, 23 Aug., 2020. URL: https://venturebeat.com/ai/the-term-ethical-ai-is-finally-starting-to-mean-something/ (accessed: Jun. 7 2023).

[6] J. S. Borg, Four investment areas for ethical AI: Transdisciplinary opportunities to close the publication-to-practice gap, Big Data & Society, vol. 8, no. 2, 205395172110401, 2021, doi: 10.1177/20539517211040197.

[7] M. Braun and P. Hummel, Data justice and data solidarity, Patterns (New York, N.Y.), vol. 3, no. 3, p. 100427, 2022, doi: 10.1016/j.patter.2021.100427.

[8] C. Burr and D. Leslie, Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies, AI Ethics, vol. 3, no. 1, pp. 73–98, 2023, doi: 10.1007/s43681-022-00178-0.

[9] G. Cachat-Rosset and A. Klarsfeld, Diversity, Equity, and Inclusion in Artificial Intelligence: An Evaluation of Guidelines, Applied Artificial Intelligence, vol. 37, no. 1, p. 2176618, 2023, doi: 10.1080/08839514.2023.2176618.

[10] A. Jobin, M. Ienca, and E. Vayena, The global landscape of AI ethics guidelines, Nat Mach Intell, vol. 1, no. 9, pp. 389–399, 2019, doi: 10.1038/s42256-019-0088-2.

[11] S. Caton and C. Haas, Fairness in Machine Learning: A Survey, ACM Computing Surveys, 2020, doi: 10.48550/arXiv.2010.04053.

[12] D. Pessach and E. Shmueli, A Review on Fairness in Machine Learning, ACM Computing Surveys, vol. 55, no. 3, pp. 1–44, 2022.

[13] J. Branford, Experiencing AI and the Relational 'Turn' in AI Ethics, International Conference on Computer Ethics, 2023.

[14] R. Binns, Fairness in Machine Learning: Lessons from Political Philosophy, in: Proceedings of Machine Learning Research, New York, 2017, pp. 1–11. URL: https://ssrn.com/abstract=3086546

[15] I. M. Young, Ed., Justice and the politics of difference. Princeton: Princeton University Press, 1990.

[16] I. M. Young, Responsibility for justice. New York, Oxford: Oxford University Press, 2011.

[17] E. Anderson, What is the Point of Equality?, Ethics, vol. 109, no. 2, pp. 287–337, 1999, doi: 10.1086/233897.

[18] R. Nath, Relational egalitarianism, Philosophy Compass, vol. 15, no. 7, 2020, doi: 10.1111/phc3.12686.

[19] Y. Kong, Are "Intersectionally Fair" AI Algorithms Really Fair to Women of Color? A Philosophical Analysis, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, 2022, pp. 485–494.

[20] N. Tomasev, K. R. McKee, J. Kay, and S. Mohamed, Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event USA, 2021, pp. 254–265.

[21] M. Andrus and S. Villeneuve, Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul Republic of Korea, 2022, pp. 1709–1721.
[22] T. Krupiy, A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective, Computer Law & Security Review, vol. 38, p. 105429, 2020, doi: 10.1016/j.clsr.2020.105429.
[23] A. L. Hoffmann, Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, Information, Communication & Society, vol. 22, no. 7, pp. 900–915, 2019, doi: 10.1080/1369118X.2019.1573912.
[24] A. Birhane, Algorithmic injustice: a relational ethics approach, Patterns (New York, N.Y.), vol. 2, no. 2, 2021, doi: 10.1016/j.patter.2021.100205.
[25] A. Gwagwa, E. Kazim, and A. Hilliard, The role of the African value of Ubuntu in global AI inclusion discourse: A normative ethics perspective, Patterns (New York, N.Y.), vol. 3, no. 4, p. 100462, 2022, doi: 10.1016/j.patter.2022.100462.
[26] Z. Tacheva, Tracking a critical look at the critical turn in data science: From “data feminism” to transnational feminist data science, Big Data & Society, vol. 9, no. 2, 205395172211129, 2022, doi: 10.1177/20539517221112901.
[27] B. Green, Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness, Philosophy & Technology, vol. 35, no. 4, 2022, doi: 10.1007/s13347-022-00584-6.
[28] C. Lu, J. Kay, and K. R. McKee, Subverting machines, fluctuating identities: Re-learning human categorization, in: FAccT ’22, 2022, pp. 1005–1015.
[29] A. Zimmermann and C. Lee-Stronach, Proceed with Caution, Canadian Journal of Philosophy, vol. 52, no. 1, pp. 6–25, 2022, doi: 10.1017/can.2021.17.
[30] Leavy, S., Siapera, E., & O'Sullivan, B. Ethical Data Curation for AI: An Approach based on Feminist Epistemology and Critical Theories of Race. In Conference on Artificial Intelligence, Ethics and Society (AIES), Virtual Event, USA, 2021
[31] Zajko, M. Conservative AI and social inequality: conceptualizing alternatives to bias through social theory. AI & SOCIETY, 36(3), 1047–1056, 2021, https://doi.org/10.1007/s00146-021-01153-9