Interdisciplinary Perspectives on the (Un)fairness of Artifcial Intelligence

Starke, C.; Blanke, T.; Helberger, N.; Smets, S.; de Vreese, C.

DOI
10.1007/s11023-025-09722-3

Publication date
2025

Document Version
Final published version

Published in
Minds and Machines

Minds and Machines (2025) 35:22
https://doi.org/10.1007/s11023-025-09722-3

EDITORIAL

Interdisciplinary Perspectives on the (Un)fairness of Artificial Intelligence

Christopher Starke[1] · Tobias Blanke[2] · Natali Helberger[3] · Sonja Smets[2] · Claes de Vreese[1]

Published online: 5 May 2025

1 Introduction

As artificial intelligence (AI) systems increasingly shape decisions across various societal domains, concerns about bias and unfairness have emerged over the last few years. Prominent cases illustrate how algorithmic decision-making can perpetuate and amplify discrimination, often to the detriment of already marginalized groups: In the judiciary, an algorithm used at US courts disproportionately classified racial minorities as high-risk recidivists (Dressel & Farid, 2018). In hiring, Amazon's algorithm disadvantaged female candidates by replicating historical patterns of bias embedded in the company's past hiring practices (Kelly-Lyth, 2021). In healthcare, AI-driven support systems systematically prioritized White patients, downplaying the needs of Black patients and granting earlier access to treatment (Obermeyer et al., 2019). But these cases are only the tip of the iceberg. Examples range widely from skin cancer detection software exhibiting lower accuracy for individuals with darker skin tones (Rezk et al., 2022) to faulty welfare algorithms wrongly accusing thousands of financially vulnerable families of fraud (Oldenhof et al., 2024).

This special issue moves beyond diagnosing AI unfairness in different domains. Bringing together insights from philosophy, social sciences, and computer science, it shifts the focus from problem analysis to solutions. The contributions explore how AI can be designed and deployed to advance fairness, offering concrete strategies to mitigate harm and harness AI for greater fairness. We thereby answer a call voiced by many in the academic research community. Since the launch of the ACM FAccT conference in 2018, the study of algorithmic (un)fairness has expanded far beyond

✉ Christopher Starke
c.d.r.o.starke@uva.nl

[1] Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, The Netherlands

[2] Institute for Logic, Language, and Computation, University of Amsterdam, Amsterdam, The Netherlands

[3] Institute for Information Law, University of Amsterdam, Amsterdam, The Netherlands

its origins in computer science. Today, it engages a truly interdisciplinary community, including philosophers, legal scholars, social scientists and the humanities, who explore both the technical and normative dimensions of fairness. This expansion reflects a growing awareness that fairness cannot be treated as a mathematical or engineering problem in computer science alone but requires a closer examination of societal structures, historical inequities, and power imbalances.

As AI systems become deeply embedded in daily life, the urgency of ensuring fairness in their design, deployment, and oversight cannot be overstated. With AI services like ChatGPT and Gemini making powerful systems more accessible, societies are at a crossroads between navigating the risks and reaping the benefits of AI: On the one hand, algorithmic biases can create a self-reinforcing cycle where existing discrimination informs algorithmic models, which in turn shape new harmful decisions. On the other hand, AI can be a new tool to detect existing inequalities and correct deeply embedded human biases, thereby contributing to more societal fairness.

2 Intricate Problems of AI Fairness

The principle of fairness is a central goal in the design of AI systems (Jobin et al., 2019). Fairness in this context requires that AI-driven decisions avoid unjust, discriminatory, or disproportionately harmful outcomes (Starke et al., 2022). Numerous formal definitions of fairness have been proposed, often drawing from foundational philosophical theories such as Rawls's difference principle (Franke, 2021, 2024), deontic egalitarianism (Holm, 2023), and the principle of equality of opportunity (Binns, 2018). Verma and Rubin (2018) identified three primary categories for these definitions. Statistical fairness metrics, such as statistical parity (Dwork et al., 2012), compare predicted probabilities, actual outcomes, and group distributions to measure equity. Similarity-based approaches, including fairness through awareness (Dwork et al., 2012), argue that individuals with similar characteristics should receive similar treatment, regardless of group classifications. Causal reasoning models, such as counterfactual fairness (Kusner et al., 2017), use structural equations to estimate the influence of sensitive attributes and design systems that minimize unjust disparities. Expanding on these frameworks, Zafar et al. (2017) proposed preference-based fairness, which prioritizes the improvement of benefits for one group compared to another.

The wide range of fairness definitions reflects diverse perspectives on what fairness entails, yet these conceptions can also conflict with one another (Kleinberg et al., 2016). Consequently, as tensions between different fairness criteria arise, scholars emphasize the need to consider the broader social context when determining which fairness standard is appropriate for a given application (Lepri et al., 2018).

AI unfairness manifests itself as disparate outcomes in the real world, disproportionately affecting certain social groups. As bias can emerge at multiple stages of the AI lifecycle, AI unfairness is hard to shake. Training datasets frequently mirror existing societal inequalities, and models trained on such data risk perpetuating those same patterns of discrimination (Ntoutsi et al., 2020). Mitigating data

bias requires meticulous attention to how data is collected and processed, including techniques such as data augmentation, re-sampling, and the application of de-biasing algorithms (Mandhala et al., 2022). However, bias in training data often arises from limited data availability for minority groups, making it a challenging issue to resolve. Simply collecting more data from these groups is insufficient, as it raises ethical concerns, including disproportionate scrutiny of marginalized populations and the large-scale collection of sensitive personal data.

Bias also arises from algorithmic design choices, such as the selection of sensitive features, objective functions, and optimization strategies, which may deliberately or inadvertently prioritize efficiency or accuracy over fairness (Ntoutsi et al., 2020). For example, in medical diagnosis optimizing for fairness may take a backseat to ensuring high overall detection accuracy. Further complicating the issue, algorithmic systems are often deployed in institutional and societal contexts already shaped by systemic inequality (Flowerman, 2023). For instance, politically motivated choices in implementing AI decision-making can lead to targeted actions against certain groups. The lack of diversity in development teams and limited attention to the perspectives of marginalized communities can reinforce these patterns, making the issue not just technical but also deeply structural.

3 Leveraging AI for Advancing Fairness

While AI has undoubtedly contributed to societal inequities, it also holds the potential to serve as a tool for advancing fairness. Scholars have demonstrated that automated decision-making, when properly designed, can outperform human decision-makers in terms of impartiality in specific cases (Lobel, 2022). Advances in debiasing techniques, such as re-weighting datasets and fairness-aware machine learning models, offer promising pathways for mitigating algorithmic harm (Bird et al., 2019). Scholars even go as far as to argue that AI fairness is ultimately an engineering problem, easier to fix than the cognitive biases deeply rooted in humans (Mullainathan, 2019). AI can actively promote societal equality and fairness by identifying and correcting discriminatory patterns. For example, targeted data collection and analysis have enabled companies to trace and prosecute workplace harassment (Lobel, 2022). AI-powered legal aid systems can improve access to justice for underprivileged communities (Han et al., 2024). AI-based anti-corruption tools have the potential to detect and expose abuses of power in public procurement processes (Köbis et al., 2022).

Much of the literature on AI fairness has focused on AI as a source of problems. Through this special issue, we join the call to explore the potential of AI to promote greater fairness in areas of society where detecting and addressing real-world injustices was previously impossible or unlikely. Achieving this requires interdisciplinary research to uncover where and how AI fairness can emerge and identify strategies for its implementation. This special issue aims to advance the literature in four key dimensions. It (1) challenges the dominant philosophical premise of algorithmic agency (see Korecki et al. in this SI), (2) analyzes human and institutional factors contributing to AI (un)fairness (see Langer et al., Szafran & Bach, and Odilla

in this SI), (3) outlines how more elaborate fairness metrics can contribute to fairer AI decisions (Bachmann et al., Purificato et al. in this SI), and (4) explores concrete examples where the implementation of AI has the potential to contribute to fairer societies (Carstens & Friess, Achterhold et al., and Kozcuer et al. in this SI).

4 Conclusion

As AI becomes increasingly ingrained in societies worldwide, it shapes access to resources, opportunities, and social outcomes. Numerous examples have shown that it can also reinforce and perpetuate existing inequalities. Yet, AI also has the potential to enhance fairness. This special issue goes beyond simply analyzing these fairness-related challenges. Instead, it aims to provide deeper insights into solutions for mitigating bias and explores how AI can be harnessed to create fairer societies. Achieving this requires collaboration across academic disciplines, as AI fairness is not just a technical issue but one that is deeply connected to societal values and ethical considerations. The contributions in this issue highlight the importance of an interdisciplinary approach, demonstrating that fairness in AI cannot be understood in isolation from broader social, cultural, and historical contexts. It is not enough to evaluate AI systems based on present-day standards; we must also consider how societal values evolve over time. By doing so, we can ensure that AI systems are designed not only to address existing biases but also to anticipate and adapt to the changing needs of society, ultimately helping to build a more just and inclusive future.

References

Binns, R. (2018). What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy, 16*(3), 73–80.

Bird, S., Kenthapadi, K., Kiciman, E., & Mitchell, M. (2019). Fairness-aware machine learning: Practical challenges and lessons learned. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. WSDM '19: The Twelfth ACM International Conference on Web Search and Data Mining, Melbourne VIC Australia. https://doi.org/10.1145/3289600.3291383

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances, 4*(1), eaao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12: Innovations in Theoretical Computer Science, Cambridge Massachusetts. https://doi.org/10.1145/2090236.2090255

Flowerman, C. H. (2023). (Some) algorithmic bias as institutional bias. *Ethics and Information Technology*, *25*(24). https://doi.org/10.1007/s10676-023-09698-7

Franke, U. (2021). Rawls's original position and algorithmic fairness. *Philosophy & Technology, 34*(4), 1803–1817.

Franke, U. (2024). Rawlsian algorithmic fairness and a missing aggregation property of the difference principle. *Philosophy & Technology, 37*(87). https://doi.org/10.1007/s13347-024-00779-z

Han, W., Shen, J., Liu, Y., Shi, Z., Xu, J., Hu, F., Chen, H., Gong, Y., Yu, X., Wang, H., Liu, Z., Yang, Y., Shi, T., & Ge, M. (2024). LegalAsst: Human-centered and AI-empowered machine to enhance court productivity and legal assistance. *Information Sciences, 679*(121052), 121052.

Holm, S. (2023). Egalitarianism and algorithmic fairness. *Philosophy & Technology, 36*(6). https://doi.org/10.1007/s13347-023-00607-w

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Kelly-Lyth, A. (2021). Challenging biased hiring algorithms. *Oxford Journal of Legal Studies, 41*(4), 899–928.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1609.05807

Köbis, N., Starke, C., & Rahwan, I. (2022). The promise and perils of using artificial intelligence to fight corruption. *Nature Machine Intelligence*, *4*, 418–424. https://www.nature.com/articles/s42256-022-00489-1

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual Fairness. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1703.06856

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611–627.

Lobel, O. (2022). *The equality machine*. PublicAffairs.

Mandhala, V. N., Bhattacharyya, D., & Midhunchakkaravarthy, D. (2022). Need of mitigating bias in the datasets using machine learning algorithms. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, 1–7.

Mullainathan, S. (2019). Biased Algorithms Are Easier to Fix Than Biased People. *The New York Times*. https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1356

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

Oldenhof, L., Kersing, M., & van Zoonen, L. (2024). Sphere transgressions in the Dutch digital welfare state: Causing harm to citizens when legal rules, ethical norms and quality procedures are lacking. *Information, Communication and Society, 27*(15), 2704–2720.

Rezk, E., Eltorki, M., & El-Dakhakhni, W. (2022). Improving skin color diversity in cancer detection: Deep learning approach. *JMIR Dermatology, 5*(3), e39143.

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society, 9*(2), 205395172211151.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. ICSE '18: 40th International Conference on Software Engineering, Gothenburg Sweden. https://doi.org/10.1145/3194770.3194776

Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P., & Weller, A. (2017). From parity to preference-based notions of fairness in classification. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/1707.00010