Integrating Fairness in AI Development: Technical Insights from the fAIr by design Framework

Mira Reisinger[1], MA, Rania Wazir[1], PhD

[1]*leiwand.ai, Vienna, Austria*

Abstract

This paper discusses the necessity of integrating fairness into the development of trustworthy AI systems, focusing on methods and tools designed within the *fAIr by design* project - a collaborative approach to guide development teams towards the creation of non-discriminatory AI systems. Practical applications, challenges, and recommendations based on real-world use cases are shared from a data science and machine learning team perspective. The paper advocates for continuous learning, diverse team assembly, and ongoing monitoring to ensure AI systems remain fair and inclusive, encompassing the whole life cycle of AI systems.

Keywords

Fairness in AI Development, Assurance Case, Use Cases, Ethical decision-making

1. Introduction

Artificial Intelligence (AI) has become increasingly integrated into various aspects of society, from communications to recruitment[1] to justice systems[2]. However, there are serious concerns regarding potential unwanted biases and discrimination embedded in AI systems. These concerns are not unfounded, as numerous examples of biased and discriminatory applications and their significant negative impacts on individuals and communities have been revealed [3, 4].

The challenge begins with defining fairness — a concept that proves elusive across disciplines. Drawing upon Mehrabi et al.[2], fairness in decision-making is ideally the absence of prejudice or favoritism towards any individual or group based on inherent or acquired characteristics. The evolving nature of fairness definitions highlights the challenge of addressing the intertwined issues of bias, fairness, and discrimination. This requires stakeholders and developers to work together to develop context-specific definitions of fairness and non-discrimination, which include acceptable thresholds and measurable metrics. Verma and Rubin[5] offer insights into defining and measuring fairness, involving various metrics such as predictive outcomes, similarity measures, or causal reasoning.

Recognizing the dynamic nature of fairness, influenced by societal norms and technological advancements, adds complexity to the effort of implementing fairness in AI systems but also enables AI systems to be finely tuned to balance fairness with performance. This balanced approach requires ongoing dialogue and adjustments between ethical principles and system efficiency, ensuring continuous alignment through monitoring and iterative enhancements.

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

mira.reisinger@leiwand.ai (M. Reisinger); rania.wazir@leiwand.ai (R. Wazir)

The research consortium *fAIr by design*[6], funded by the Austrian National Foundation for Research, Technology and Development (FFG)[7], has devised a framework aimed at aiding organizations in incorporating fairness requirements into their development processes. This paper reflects on the utilization of the tools and methods[8, 1] introduced to 3 AI system development teams to ensure adherence to fairness and trustworthiness principles.

2. Leveraging Assurance Cases for Fairness in AI Systems

This chapter delves into a strategy employed by the *fAIr by design* team to navigate the challenges of developing algorithms with embedded fairness constraints. Drawing on the adapted Assurance Case method [8], the strategy includes a comprehensive discussion about the purpose and role of an AI system, the analysis and identification of potential fairness and discrimination risks followed by thorough testing for those risks, and adoption of mitigation strategies.

2.1. The Assurance Case method (AC)

Originating from safety engineering, the Assurance Case method offers a structured pathway for translating high-level goals into specific, actionable, and verifiable technical specifications [9, 10]. It is a collaborative approach fostering a holistic perspective on fairness in AI development, promoting the creation of ethically sound and socially responsible systems. The AC facilitates close cooperation among social science, data science, and start-up development teams in the *fAIr by design* project, serving as a valuable tool for integrating fairness into AI systems.

Components of the method include claims, sub-claims, evidence and reasoning, which enable the project team to identify and systematically integrate a comprehensive understanding of the AI system and fairness needs through the definition of sub-claims, each supported by (technical) evidences. The overall process included developing evidences and tests, a thorough exploration of model selections, fairness metrics, and the application of mitigation strategies.

The AC emphasizes the crucial role of integrating "challengers" (technical or non-technical) into the AI development process to identify and point out aspects that are not clear and can lead to unfairness. Having a comprehensive understanding of each component of the AI system from the outset is invaluable, and the involvement of "challengers" enhances this understanding by ensuring that fairness considerations are addressed. The "challenger" can assist in clarifying data needs, testing protocols, evaluation metrics, and mitigation strategies, laying the groundwork for in-depth fairness testing. This also equips the team to manage potential future challenges.

3. Practical Applications and Use Cases

This chapter reports on the application of the methodologies and tools found during *fAIr by design.* The implementation of the AC was done together with social scientists and legal experts, but learnings and recommendations center on technical application and data science perspective.

3.1. Learnings and Recommendations

At the onset of developing a fair AI system, it's important from a technical and social-science standpoint to engage in a series of critical inquiries and clarifications with the use case partner, which include:

- The definition of fairness from the partner's perspective.
- The characteristics that constitute a fair AI system for them.
- The (envisioned) structure of the AI system and it's components.
- The risks identified and the sub-claims that can be substantiated with technical evidence.
- The feasible tests, along with a potential responsible person.
- The prerequisites for conducting these tests, including necessary data, knowledge, and resources.
- The applicable metrics and thresholds for fairness.
- The mitigation strategies to be employed should these thresholds not be met.

3.1.1. Tools and Resources

The use of methods and tools developed within the Assurance Case or those available as open access resources[11, 12], such as the AI canvas and ethics checklists, are particularly recommended. The Data Science Ethics Checklist from deon.org[12], applied as an iterative process, has been shown to significantly support the technical progress. If well done it can be part of the documentation requirements of high risk AI systems[13]. The checklist helps clarifying the data science and machine learning maturity level of the partners, and aligns well with the need for clarity and efficacy common in development teams[1].

3.1.2. Assessing and Building Knowledge

All steps of fairness testing require data quality and overall machine learning and data science expertise (e.g. ablation studies, hyper-parameter optimization or inverse relation modeling). Addressing biases in training and evaluation data is vital to prevent AI systems from replicating or exacerbating existing inequities. For some companies it can be a challenge to build the necessary knowledge around data quality, conducting systematic evaluations, and testing model components, which should be addressed from the onset.

3.1.3. Assemble a Diverse and Competent Team

Building a proficient team that specializes in appropriate data science and machine learning methods is essential to circumvent fairness testing pitfalls effectively. It requires a concerted effort to integrate knowledge on fairness into AI, delineating clear responsibilities among senior management and development teams. Fairness, much like other critical quality criteria, must be integrated into a wide array of business processes, gaining prominence especially in high-risk AI systems[13]. The establishment of interdisciplinary teams is key to facilitating in-depth discussions and making informed decisions regarding fairness.

3.1.4. Continuous Validation and Monitoring

The establishment of regular auditing and accountability mechanisms is pivotal in upholding non-discriminatory practices in AI. Continuous monitoring and evaluation of AI systems enable organizations to proactively identify and address any potential biases or discriminatory outcomes, thereby demonstrating their commitment to fairness and ethical AI development.

3.1.5. You think you know what fairness means – until you ask others

Another critical aspect of promoting non-discriminatory AI development is fostering collaboration and transparency. Organizations should actively seek partnerships with diverse stakeholders, including ethicists, community representatives, and regulatory bodies, to gain insights into the potential biases and discriminatory risks in AI systems. By promoting transparency in AI development processes, organizations can build trust and accountability with the public, while ensuring that fairness remains an integral part of the AI life cycle. This is useful as a quality assurance tool, ensuring that the product being developed actually satisfies market needs.

3.2. Challenges and Limitations

Collaborative efforts with partners have provided valuable insights into the practical challenges and strategies for developing non-discriminatory AI systems. These collaborative endeavors shed light on the significance of data science and machine learning maturity, as well as the willingness to invest time, effort, and resources into crafting fair, non-discriminatory, and trustworthy AI systems. However, navigating fairness throughout the AI life cycle presents challenges, including the dynamic nature of fairness definitions and the intricacies of measuring fairness.

The *fAIr by design* use cases, focusing on small companies (SMEs), start-ups and cultural organizations/NGOs - each with distinct applications, domains, and developmental stages - underscore the diverse landscape within which fairness considerations are embedded, emphasizing the importance of contextual understanding and adaptability. However, we have not been able to work with larger, more established organizations, with a more advanced data management structure, but may face other challenges.

4. Conclusion

The journey towards non-discriminatory AI development requires the concerted efforts of organizations, policymakers, and society as a whole. *fAIr by design*, including the Assurance Case, provides a structured approach to the development of fair AI systems. Recognizing the contextual nature of fairness, the adoption of ethical AI principles should be accompanied by continuous training for development teams, enabling them to incorporate ethical guidelines into their day-to-day decision-making processes. In this project we have seen how beneficial it is for social sciences, data sciences and use case partners working together. Moving forward, the exploration of cross-industry collaborations and the proposition of structured frameworks for engaging organizations in fairness discussions are promising directions for the advancement

of non-discriminatory AI. The potential to draw on diverse perspectives and expertise, could ultimately lead to industry-wide standardized approaches, thus bringing us closer to a future where ethical principles in AI are upheld across the board.

References

[1] S. Cepeda, L. Kunze, G. Leimüller, L. Müller-Kress, M. Stöger, R. Wazir, Requirements for developing fair ai systems, https://www.fairbydesign.eu/publications, 2022. [Accessed 29-03-2024].

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3457607. doi:10.1145/3457607.

[3] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown Publishing Group, New York, NY, USA, 2016.

[4] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, Science Advances 4 (2018) eaao5580. doi:10.1126/sciadv.aao5580.

[5] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness, FairWare '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–7. doi:10.1145/3194770.3194776.

[6] fair by design, https://www.fairbydesign.eu, 2024. [Accessed 29-03-2024].

[7] fAIr by design - Entwicklung eines neuen Prozessmodells für nicht-diskriminierende Künstlicher Intelligenz | FFG, https://www.ffg.at/content/fair-design-entwicklung-eines-neuen-prozessmodells-fuer-nicht-diskriminierende-kuenstlicher, 2021. [Accessed 29-03-2024].

[8] L. Kunze, G. Leimüller, L. Müller-Kress, M. P. Hauer, Method handbook: Assurance cases for fair ai systems, https://www.fairbydesign.eu/publications, 2024. [Accessed 29-03-2024].

[9] M. P. Hauer, L. Müller-Kress, G. Leimüller, K. Zweig, Using assurance cases to assure the fulfillment of non-functional requirements of AI-based systems - lessons learned, in: 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), IEEE, 2023.

[10] M. P. Hauer, R. Adler, K. Zweig, Assuring fairness of algorithmic decision making, in: 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), 2021, pp. 110–113. doi:10.1109/ICSTW52544.2021.00029.

[11] J. Unadkat, What is Test Driven Development (TDD)?, https://www.browserstack.com/guide/what-is-test-driven-development, 2023. [Accessed 29-03-2024].

[12] deon, Data science ethics checklist, https://deon.drivendata.org/, 2019. [Accessed 29-03-2024].

[13] European Parliament and Council of European Union, Regulation (EU) no 2024/1689 Annex III. High-risk AI systems referred to in Article 6(2), https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024.