RRI as Opportunity for Enhancing Procedural Fairness in Machine Learning

Keywords: Algorithmic Fairness, XAI, RRI, Participation, Procedural Fairness

Extended Abstract

Machine Learning (ML), and in particular automated decision-making (ADM) systems are nowadays utilized for various application fields. Growing in their impact on life-influencing high-stakes decisions, these systems can reinforce already existing biases, produce unfair or discriminatory outcomes, or systematically reinforce stereotypes and inequalities [1]. Consequently, the fair ML community made considerable efforts to systematically and technically address fairness in ADM applications. This has led to over twenty different metrics that all in their way claim to measure fairness, but also introduce contradictions between them [2, 3]. Their common ground is to try to quantify fairness by mathematical methods aiming to analyze the outcome of a decision in terms of its *distributive* justice [4]. Rising attention is devoted to the challenges and consequences of ill-defined measurements and their limited embedding in social theories [5]. To reveal discriminatory patterns, an increasing call for explainable and transparent algorithms has risen (e.g., [6]). Methods of Explainable AI (XAI) should enable users to identify and mitigate biases and enhance algorithmic fairness. In both fields of research – XAI, fairness of AI, and their interplay –, various empirical studies on people's perception of the respective concepts have been conducted (e.g., [7]). A lack of consensus regarding the 'best' approach to ensure algorithmic fairness is however observable in both the social and technical research. From a normative perspective, this is not too surprising, as subjectivity lies in the nature of fairness: in ethical discussions, fairness is defined as the individual moral evaluation of rules of conduct [8]. In contrast, justice refers to a standard of rightness, leading to impartial and consistent rules of conduct [9]. In the far-reaching debate of moral and political philosophers, they made a significant effort to elaborate on the concept of justice, figuring out that *distributive* questions are only one dimension in the complex conception of justice [9]. However, this perspective falls rather short in the current research on algorithmic fairness [10].

We argue that algorithmic justice should be considered more holistically in the attempt of ensuring algorithmic fairness and that procedural fairness should take the center stage in the research attempts. We further argue that a promising approach in doing so may be reached by civil participation and public engagement[1] as formulated within the framework of Responsible Research and Innovation (RRI) – a concept aiming for a better alignment of research and innovation with societal needs [11, 12]. Within the RRI framework, five policy agendas were formulated, among them the promotion of public engagement, i.e., to involve diverse stakeholders and civil society in the whole process of research and development. Participatory approaches have been widely adopted in human-computer interaction and engineering but are seldom in the design of algorithmic applications. However, we argue that in the context of fair ADM, public engagement has the potential to address several issues that have been risen by the research of previous years.

[1] We distinguish engagement as top-down initiative from governments and participation as bottom-up initiative by citizens.

First of all, it may be required by a moral intuition: those affected by ADM (decision-subjects) have a legit demand to understand the decision and consequences. As shown by XAI research, the requirements to enhance comprehensibility differ between experts and non-experts [13]. Public engagement can foster the identification of different needs and enable those affected to contest the decision. This may also serve as one step toward reducing the responsibility gap in algorithmic systems [14, 15]. Meeting the knowledge responsibility of ADM users, we must ask: What should the user of technology have known? Which information would he/she have needed beforehand to secure a socially responsible application of AI? Thus, public engagement benefits not only the ADM subjects but also its users. And also, those, who develop ADM may benefit from a reverse learning process, by including more broad knowledge in the development of technology by using available knowledge from society. Currently, too many far-reaching high-stakes decisions are made by a small coding elite [16]. Finally, by considering society's needs and values and involving diverse stakeholders in the process through engagement, more inclusive technology may be secured. Including the public in the process may also enhance awareness of the far-reaching consequences of ADM and thus public engagement may function to train participants.

In our contribution, we illustrate why RRI, and especially the perspective of public engagement, can and should enhance the debate on algorithmic fairness. We aim to demonstrate several reasons to foster participatory research and public engagement as formulated within the RRI concept in the context of fair ADM and promote future research in the field of procedural fairness in ADM.

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning*: fairmlbook.org, 2022.

[2] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," 2017, doi: 10.48550/arXiv.1609.05807.

[3] A. Narayanan, "21 fairness definitions and their politics," Mar. 1 2018. [Online]. Available: https://www.youtube.com/watch?v=jIXIuYdnyyk

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

[5] C. Wagner, M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor, and T. Eliassi-Rad, "Measuring algorithmically infused societies," *Nature*, vol. 595, no. 7866, pp. 197–204, 2021, doi: 10.1038/s41586-021-03666-1.

[6] European Commission and Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*: Publications Office, 2019.

[7] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human Perceptions of Fairness in Algorithmic Decision Making," in *Proceedings of the 2018 World Wide Web Conference*, Lyon, France, 2018, pp. 903–912.

[8] B. Goldman and R. Cropanzano, ""Justice" and "fairness" are not the same thing," *J. Organiz. Behav.*, vol. 36, no. 2, pp. 313–318, 2015, doi: 10.1002/job.1956.

[9] D. Miller, *Justice*, 2017. Accessed: Feb. 14 2023. [Online]. Available: https://plato.stanford.edu/entries/justice/

[10] 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*, 2022. Accessed: Jan. 26 2023.

[11] R. Owen, P. Macnaghten, and J. Stilgoe, "Responsible research and innovation: From science in society to science for society, with society," *Science and Public Policy*, vol. 39, no. 6, pp. 751–760, 2012, doi: 10.1093/scipol/scs093.

[12] R. von Schomberg, *A Vision of Responsible Research and Innovation*: John Wiley & Sons, Ltd, 2013.

[13] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," in *Conference on Fairness, Accountability, and Transparency (FAT* '19)*, FAT* '19, Ed., Atlanta, GA, USA. ACM, New York, NY, USA, 2019, pp. 279–288.

[14] A. Adensamer, R. Gsenger, and L. D. Klausner, ""Computer says no": Algorithmic decision support and organisational responsibility," *Journal of Responsible Technology*, 7-8, p. 100014, 2021, doi: 10.1016/j.jrt.2021.100014.

[15] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics Inf Technol*, vol. 6, no. 3, pp. 175–183, 2004, doi: 10.1007/s10676-004-3422-1.

[16] C. D'Ignazio, *DATA FEMINISM*. [S.l.]: MIT Press, 2023.