HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230118

Functional Requirements for Interactive Bias-Audit Tools

Daphne LENDERS [a,1] and Toon CALDERS [a]
[a] *University of Antwerp, Campus Middelheim*
ORCiD ID: Daphne Lenders https://orcid.org/0000-0002-9839-9077, Toon Calders https://orcid.org/0000-0002-4943-6978

Abstract. Interactive toolkits can be used to audit automated decision-making models for discriminatory biases. In this paper, we present a rubric listing the functional requirements such toolkits need to fulfil to make them usable in practice. The rubric was based on a literature review of interview studies with industry practitioners, and other possible auditors.

Keywords. Fair ML, Bias Audit, Tools for Audits, Interactive Tools

Over the past years, more and more automated decision-making (ADM) models are being developed for tasks usually performed by humans, such as credit scoring or job recruitment. Though powerful and potentially time-saving, these models come with the risk of mirroring and amplifying discriminatory patterns recorded in the data they are based on. In other words, their decisions might (directly or indirectly) favour some groups of the population, while disfavouring others, based on sex, age, race, or other characteristics of the decision subjects [10]. In response to growing concerns about the unfairness of ML, research organizations and legal institutions have called out the importance of auditing the bias of ADM systems before they are deployed, a process that in the case of hiring and employment systems has already become mandatory in New York City [8] and for which regulations are being proposed in the EU, in the form of the EU AI Act [4]. Conducting an audit is a lengthy and complex process in which a lot of considerations need to be made, like how to define bias, which groups to include in the audit, and to which extent some disparate treatments of groups might be justifiable. To streamline and simplify this process, interactive tools can help, which let auditors inspect a model's predictions and input data in a visual way [1,2,13,14]. As more of these tools are being developed, all of them differing in the functionalities they provide, the question arises of which functionality they actually need to offer to make them usable in realistic settings. In an attempt to answer this question, interview studies with potential auditors have been conducted, in which essential design considerations and requirements were identified [3,5,6,7,9,11,12]. Still, an extensive overview of these requirements is lacking, which is why we conducted a literature review to compile a rubric that brings the results of these studies together. The resulting rubric is presented in Figure 1. As can be seen, the first part of the rubric relates to a tool's functionality to let users inspect the predictions of an ADM model for possible biases. In other words, tools should help users in

[1] Corresponding Author

examining whether the models' outputs or errors disfavour some groups or individuals of the population. Relating to this, it is among others important that tools do not merely support bias analysis on one binary sensitive attribute (e.g., race: white vs. non white) but enable bias analysis in more complex settings, when intersectional groups might be affected or when access to sensitive attributes is missing.

The second family of requirements relates to tools' functionality to let users inspect the input data. This inspection is crucial to understand where biases in the models' predictions come from and how they can be mitigated. For instance, users need to be able to access the training labels of the data to understand whether any disparities in predictions (e.g., women being denied a loan more often than men) are also reflected there. Another example is the functionality to inspect differences in sample sizes, so that users understand why a model might have worse performance on certain population groups than on others.

The last family of requirements relates to whether tools make bias audits scalable. For instance, tools should report the confidence intervals of their bias measures so that auditors can focus on biases that reflect systemic issues of a model, rather than wasting time on "one-off" mistakes that are expected by chance. At the same time, tools should automatically highlight subgroups and individual instances that might be subject to discriminatory bias, so that an auditor does not miss important biases.

In our poster presentation, we will examine each requirement in more detail, to highlight the considerations that should go into the fairness assessment of an ADM model and help developers of future tools understand practitioners' functional needs in them.

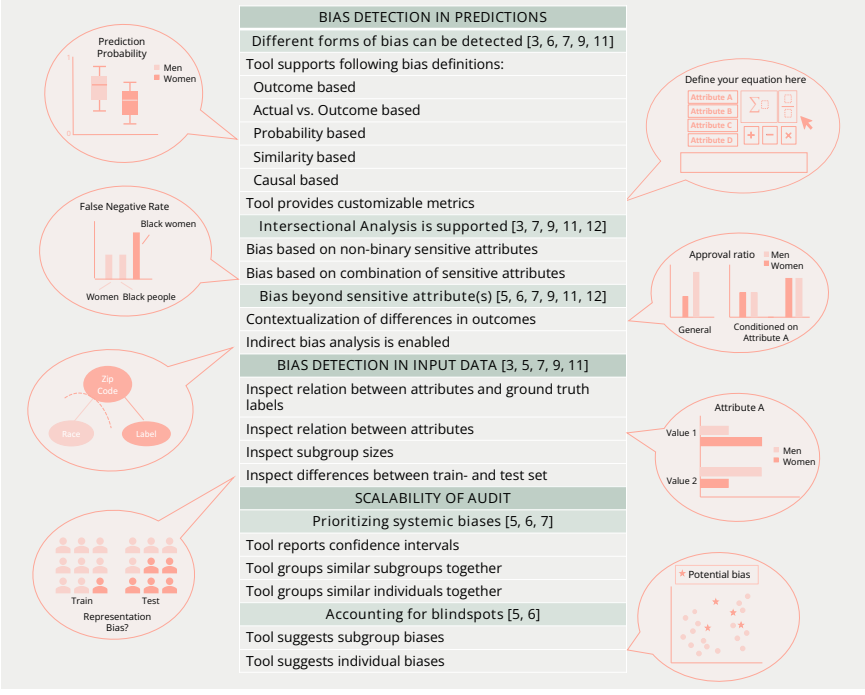

Figure 1. Rubric of requirements in interactive auditing tools. The rubric was compiled based on a literature review of the studies [3,5,6,7,9,11,12]

References

[1] Ahn, Y., & Lin, Y. R. (2019). Fairsight: Visual analytics for fairness in decision making. IEEE transactions on visualization and computer graphics, 26(1), 1086-1095.

[2] Cabrera, Á. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019, October). FairVis: Visual analytics for discovering intersectional bias in machine learning. In 2019 IEEE Conference on Visual Analytics Science and Technology (VAST) (pp. 46-56). IEEE.

[3] Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1571-1583).

[4] European Commission (2021). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206

[5] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019, May). Improving fairness in machine learning systems: What do industry practitioners need?. In Proceedings of the 2019 CHI conference on human factors in computing systems (pp. 1-16).

[6] Law, P. M., Malik, S., Du, F., & Sinha, M. (2020). Designing Tools for Semi-Automated Detection of Machine Learning Biases: An Interview Study. arXiv preprint arXiv:2003.07680.

[7] Lee, M. S. A., & Singh, J. (2021). The landscape and gaps in open source fairness toolkits. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-13).

[8] New York City Council (2021). A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools. URL: https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=Advanced&Search

[9] Nakao, Y., Strappelli, L., Stumpf, S., Naseer, A., Regoli, D., & Gamba, G. D. (2022). Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. International Journal of Human–Computer Interaction, 1-27.

[10] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... & Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1356.

[11] Richardson, B., Garcia-Gathright, J., Way, S. F., Thom, J., & Cramer, H. (2021). Towards fairness in practice: A practitioner-oriented rubric for evaluating Fair ML Toolkits. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

[12] Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 chi conference on human factors in computing systems (pp. 1-14).

[13] Wang, Q., Xu, Z., Chen, Z., Wang, Y., Liu, S., & Qu, H. (2020). Visual analysis of discrimination in machine learning. IEEE Transactions on Visualization and Computer Graphics, 27(2), (pp. 1470-1480).

[14] Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2019). The what-if tool: Interactive probing of machine learning models. IEEE transactions on visualization and computer graphics, 26(1), 56-65.