# Inside Fairness Tools: What Academic Practitioners Really Experience

Sadia Afrin Mim*, Brittany Johnson †
*†Department of Computer Science, George Mason University
*safrinmi@gmu.edu, †johnsonb@gmu.edu,

***Abstract*—Fairness in AI models has become essential as our society increasingly becoming more dependent on AI. A biased model can have a harmful impact on marginalized communities. To address this issue, practitioners have developed fairness tools over time. To understand their practical implication, we designed an interview to curate experiences with fairness tools from academic practioners'. In this paper, we discuss insights from our first round of interviews with practitioners from academia. Although numerous fairness tools have been developed, only a few industry-developed (e.g., AIF360 and Fairlearn) are practically usable and commonly employed by practitioners due to regular maintenance & visibility. The existing toolkit landscape is primarily equipped to solely handle textual data and lacks sufficient resources for language models. Our findings thus far provide insights into one perspective on fairness tool engagement; our future efforts will investigate experiences and perspectives on fairness tool support beyond traditional models.**

***Index Terms*—fairness tools, machine learning, usability**

## I. Content And Claims

To address biases and associated risks in AI models, practitioners have introduced fairness tools [1]. They have developed metrics to detect and measure bias in models and established methodologies for defining fairness [2]. Additionally, efforts have been made to mitigate bias at different stages of model training. Fairness tools promote fair alternatives to existing algorithms ensuring accountability & transparency [3].

Several research efforts have explored the availability of fairness tools [4]. Some studies have conducted human-centric evaluations to compare these tools [1], while others have examined how comprehensively these tools establish the broader concept of ethics [5]. However, as the landscape of fairness tools continues to expand, tracking their scope becomes increasingly challenging. Despite extensive research on fairness tools, we found prior works mostly examining their role within the industry practitioners. This study addresses an overlooked issue through incorporating the viewpoints of academic practitioners, whereas prior studies have mainly concentrated on industry practitioners to comprehend the adoption of fairness tools.

*The goal of our research to better understand academic practitioners' engagement with fairness tools.* To this end, we designed an interview study to elicit experiences from practitioners on the presence and use of fairness tools. In this paper, we provide insights into the first phase of our efforts where we interviewed 6 academic practitioners (some of them have industry background) regarding their interactions with fairness tools and challenges they encounter. Participants expressed varied opinions about the utility of fairness tools. Some participants found these tools helpful, as they provided essential metrics, saving them the effort of implementing these from scratch. However, others noted that the tools were limited to binary classification models and lacked applicability to a broader range of models (e.g.,Language models) & non textual dataset. Additionally, while some found the resources provided by these tools enriching and useful for beginners, others criticized the tools being not suitable for novices for lacking sufficient explainability in resources requiring regular consultation of the documentation for each operation. As we continue these efforts, we will interview practitioners to curate more detailed insight into their experiences and needs. Our research will provide insights into the ways in which we can facilitate more consistent use of fairness tools among practitioners.

## II. User Study

Before conducting interviews, we followed a two-step participant selection process, starting with a survey to assess familiarity with fairness tools and computing background, yielding 178 responses. From this survey, we randomly selected 10 individuals and used a pre-assessment form to finalize participants, leading to four interviews, plus two more via snowball sampling. In total, we interviewed six academic practitioners, two of whom also had industry experience. Each 30-minute interview focused on participants' personal experiences with fairness tools rather than empirical evaluations.

We are conducting our user study for answering the following research questions:

- **RQ1** *In what ways academic practitioners are engaging with these tools?*
- **RQ2** *What are the challenges practitioners facing while using these tools?*

Our interview script explored several key aspects of fairness tools to understand their application in academic environments. To answer *RQ1* we analyzed 1. The specific fairness tools used and application contexts, 2. Initial factors influencing tool selection, 3. The metrics required for research and their alignment with objectives, 4. The level of community support and documentation quality, 5. The user-friendliness of the tool interfaces, and 6. Participants' suggestions for tool

improvements. To address *RQ2*, we concentrated on the same factors as *RQ1*, examining whether participants encountered any challenges related to these factors.

Following the completion of our interviews, we employed a qualitative coding methodology to derive insights from our data. We began by transcribing the recorded data and then followed the lookup summarization approach to extract and directly map data to each specific research question [6]. We analyzed them following pattern coding by summarizing common themes from the interviews using notes [7]. Our participants reported using IBM AIF360, Microsoft Fairlearn, Seldonian toolkit, and Google's What-if and LIT tools in their workflow.

Participants reported using fairness tools to address biases in datasets and models across various domains including classifiers, word embeddings, and language models. The motivations ranged from enhancing fairness metrics compliance to developing new tool features and visualizations. While some focused on mitigating biases in commonly used or their own prepared dataset for machine learning models, others aimed to tackle socio-cultural biases in language models. Some found these tools as valuable as a starting point for understanding AI fairness. Our study revealed that none of the participants could depend solely on a single tool to ensure fairness in their projects which supports multifaceted nature of AI models [8]. Also participants prefer to modify existing tools to fit their needs rather than seeking another tool. Some participants prefers implementing their own metrics instead of using existing fairness tools. In future work, we can elaborate on the reasons behind this preference.

Common metrics across the respondents include Equal Opportunity [9], and Equalized Odds [10]. The utilization of other metrics like Statistical Parity, Demographic Parity, and Disparate Impact emphasizes the need for larger group-level equity, making sure that decisions don't disproportionately help or hurt any specific group [11]. In contrast, metrics like Predictive Equality and Conditional Demographic Disparity focuses fairness of specific error-prone situations. For engaging with a broader of set metrics, participants found both AIF360 and Fairlearn to be equally useful. Our analysis also revealed that participants working with language models existing fairness tool metrics are often not applicable.

Our participants offered varied feedback on the documentation of fairness tools, with most expressing some level of satisfaction. Some participants appreciated the thorough and well-developed documentation provided for tools created by major tech companies, such as AIF360, Fairlearn, What-If, and LIT. They highlighted another issue, noting that prior to using these tools, they engaged with several fairness tools with poor or non-existent documentation. For example, participants faced challenges with the documentation of academically developed tools such as the Seldonian toolkit, citing it as outdated and challenging to comprehend for those outside the development team. Furthermore, they noted that the available sample code and demonstrations are likely inadequate for accommodating the varied needs of different fairness practitioners.

Participants reported lack of beginner-friendly explanations regarding the underlying mechanisms, configuration options, and appropriate use of protected attributes. They also noted that the sample code and project examples were insufficient, and clearer guidance on fairness metrics was needed. Participants also found it difficult to start working with fairness tools because of occasional bugs, which were particularly confusing for those unfamiliar with the tool's inner workings—especially when dealing with more advanced algorithms. Furthermore, they face issues integrating custom-built datasets, as the tools do not provide clear instructions or support for adapting workflows to user-specific data inputs.

Some participants reported familiarity with the development communities behind certain fairness tools, which allowed them to provide feedback and report issues directly to the teams. In majority cases, they received responses within a few weeks. Their experience was further enhanced by the availability of interactive documentation, which, alongside their connection to the development teams, helped them navigate and utilize the tools more effectively. In contrast, others without familiarity to the team, sought help by posting on the GitHub page but did not receive any response.

Our findings reveal several noteworthy insights, many of which align with conclusions drawn in previous studies [12], [13]. Consistent findings with earlier research indicates limited effort and maintenance surrounding these tools, reinforcing claims that their typical lifespan is only 2–3 years from prior work [14]. Our current work offers additional perspective on the underlying causes of low community engagement. To enhance adoption and sustainability, expanding these tools beyond niche communities and providing more comprehensive documentation and resources could play a crucial role. Our results suggest that academic practitioners tend to prioritize the extensibility and explainability of these tools, whereas industry practitioners place greater emphasis on stability and ease of integration to product. Academic researchers also highlighted the challenges of collaborating on these projects due to their development within close-knit communities.

## III. Relevance

The main focus of VL/HCC is around human-centric computation. Therefore, designing & upgrading tools & technologies in a more usable way based on user studies can be included in this category. Prior works at VL/HCC has consistently emphasized user-centric considerations in the development and improvement of various tools, as seen in publications [4], [15], [16]. Our work analyzes user feedback on tools designed to ensure fairness and equity in AI models which aligns with the conference's objectives.

## IV. Presentation

We plan to present our findings through both a physical and virtual poster to offer flexibility and encourage audience engagement and feedback. Our poster will highlight real-world interactions between academic practitioners and fairness tools.

## REFERENCES


[1] M. S. A. Lee and J. Singh, "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, 2021.

[2] Y. Brun and A. Meliou, "Software fairness," in *Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pp. 754–759, 2018.

[3] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges," *Philosophy & Technology*, vol. 31, no. 4, pp. 611–627, 2018.

[4] S. A. Mim, J. Smith, and B. Johnson, "A taxonomy of machine learning fairness tool specifications, features and workflows," in *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 222–225, 2023.

[5] R. Y. Wong, M. A. Madaio, and N. Merrill, "Seeing like a toolkit: How toolkits envision the work of ai ethics," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, pp. 1–27, 2023.

[6] V. Braun and V. Clarke, "Thematic analysis," in *Encyclopedia of quality of life and well-being research*, pp. 7187–7193, Springer, 2024.

[7] J. Saldaña, "The coding manual for qualitative researchers," 2021.

[8] J. Franse, V. Misheva, and D. S. Vale, "Practical and open source best practices for ethical machine learning," in *Towards Trustworthy Artificial Intelligent Systems*, pp. 77–84, Springer, 2022.

[9] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira, *et al.*, "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big data and cognitive computing*, vol. 7, no. 1, p. 15, 2023.

[10] P. Awasthi, M. Kleindessner, and J. Morgenstern, "Equalized odds postprocessing under imperfect group information," in *International conference on artificial intelligence and statistics*, pp. 1770–1780, PMLR, 2020.

[11] P. Garg, J. Villasenor, and V. Foggo, "Fairness metrics: A comparative analysis," in *2020 IEEE international conference on big data (Big Data)*, pp. 3662–3666, IEEE, 2020.

[12] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 3, p. e1452, 2022.

[13] B. Richardson, J. Garcia-Gathright, S. F. Way, J. Thom, and H. Cramer, "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, 2021.

[14] S. A. Mim, F. Vares, A. Meenly, and B. Johnson, "An investigation into open source fairness tool sustainability," in *Proceedings of the 1st International Workshop on Responsible Software Engineering*, pp. 21–28, 2025.

[15] A. Alaboudi and T. D. LaToza, "An exploratory study of live-streamed programming," in *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 5–13, IEEE, 2019.

[16] L. Costa, S. Barbosa, and J. Cunha, "Programmer user studies: Supporting tools & features," in *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 163–167, IEEE, 2024.