Algorithmic Fairness Through the Lens of Metrics and Evaluation (AFME) 2024

Miriam Rateike MIRIAM.RATEIKE@IBM.COM
IBM Research Africa, Saarland University

Awa Dieng AWADIENG@GOOGLE.COM
Google DeepMind, Mila

Jamelle Watson-Daniels WATSONDANIELS@META.COM
Meta

Ferdinando Fioretto FIORETTO@VIRGINIA.EDU
University of Virginia

Golnoosh Farnadi FARNADIG@MILA.QUEBEC
McGill University, Mila

Editors: Miriam Rateike, Awa Dieng, Jamelle Watson-Daniels, Ferdinando Fioretto, Golnoosh Farnadi

1. Introduction

The discussion on defining and measuring algorithmic (un)fairness has predominantly been a focus in the early stages of algorithmic fairness research (Dwork et al., 2012; Zemel et al., 2013; Hardt et al., 2016; Zafar et al., 2017; Agarwal et al., 2018) resulting in four main fairness denominations: individual or group (Binns, 2020), statistical or causal (Makhlouf et al., 2024), equalizing or non-equalizing (Diana et al., 2021), and temporal or non-temporal fairness (Rateike, 2024). Since, much work in the field had been dedicated to providing methodological advances within each denomination and understanding various trade-offs between fairness metrics (Binns, 2020; Heidari et al., 2019; Kleinberg et al., 2017). However, given the changing machine learning landscape, with both increasing global applications and the emergence of large generative models, the question of understanding and defining what constitutes "fairness" in these systems has become paramount again.

On one hand, definitions of algorithmic fairness are being critically examined regarding the historical and cultural values they encode (Asiedu et al., 2024; Arora et al., 2023; Bhatt et al., 2022). The mathematical conceptualization of these definitions and their operationalization through satisfying statistical parities has also raised criticism of not taking into account the context within which these systems are deployed (Weinberg, 2022; Green and Hu, 2018). On another hand, it is still unclear how to reconcile standard fairness metrics and evaluations developed mainly for prediction and classification tasks with large generative models. While some works proposed adapting existing fairness metrics, e.g., to large language models (LLMS) (Li et al., 2023; Zhang et al., 2023; Gallegos et al., 2024), questions remain on how to *systematically* measure fairness for textual outputs, or even multimodal generative models (Schmitz et al., 2022; Chen et al., 2024; Lum et al., 2024). Large

generative models also pose new challenges to fairness evaluation with recent work showcasing how biases towards specific tokens in LLMs can influence fairness assessments during evaluation (Ding et al., 2024). Finally, regulatory requirements introduce new challenges in defining, selecting, and assessing algorithmic fairness (Deck et al., 2024; Laux et al., 2024; Hellman, 2020).

Given these critical and timely considerations, this workshop aimed to investigate how to define and evaluate (un)fairness in today's machine learning landscape. We were particularly interested in addressing open questions in the field, such as:

- Through a retrospective lens, what are the strengths and limitations of existing fairness metrics?
- How can we operationalize contextual definitions of fairness in diverse deployment domains?
- Given the plethora of use-cases, how can we systematically evaluate fairness and bias in large generative models?
- How do recent regulatory efforts demand the utilization of fairness metrics and evaluation techniques, and do existing ones comply with regulations?

2. Workshop

The AFME workshop was held at NeurIPS in Vancouver, Canada, on 14 December 2024. Invited and contributed talks, except one of each, as well as the roundtables and panel, were presented in person. To accommodate a larger audience, we offered two poster sessions, one in the morning and one in the afternoon. All talks and the panel were livestreamed, and all accepted papers were able to pre-record a 3-minute video available on the NeurIPS website for the registered audience.

2.1. Program

AFME 2024 featured invited talks by Kush Varshney (IBM Research), Sanmi Koyejo and Angelina Wang (Standford University), Hoda Heidari (Carnegie Mellon University) and Seth Lazar (Australian National University), five spotlight talks from authors of accepted papers, an interdisciplinary panel discussion with Sanmi Koyejo (Standford University), Hoda Heidari (Carnegie Mellon University), Seth Lazar (Australian National University), and Jessica Schrouff (Google DeepMind), and three poster sessions. In addition, we hosted roundtables consisting of discussions between invited researchers of mixed seniority and workshop participants. More than 100 individuals participated in these 1-hour sessions, which covered the following topics:

- Metrics. Invited researchers: Angelina Wang (Standford University).
- Evaluation. Invited researchers: Candace Ross (Meta AI, FAIR), Tom Hartvigsen (University of Virginia).

2.2. Contributed Papers and Extended Abstracts

AFME had two tracks: a *Paper Track* which called for 4-9 page manuscripts of novel work and an *Abstract Track*track which called for 1-page extended abstracts. We received 45 viable papers submissions and 13 extended abstracts, which were sent for double blind peer reviewing. All submissions received at least 3 reviews and on average 3.69 reviews, which led to the acceptance of 29 papers (acceptance rate w/o desk reject: 64.4%) and 9 abstracts (acceptance rate: 69.2%). Among the accepted papers, 7 works focused on fairness metrics, 15 papers were predominantly related to fairness evaluation methods, and 7 studied general fairness methods or applications; among the accepted abstracts, 3 works focused on fairness metrics, 2 on fairness evaluation methods, and 4 on general fairness methods or applications.

Among the accepted papers, 9 papers were considered for inclusion in the Proceedings, with the authors of 7 works choosing to do so. All accepted works were presented as posters during the conference, and contributions in the *Paper Track* were able to pre-record 3-minute video summaries which were available on the virtual NeurIPS website. The Program Committee included 84 reviewers for the *Paper Track* and 26 for the *Abstract Track*. Since some reviewers participated in both tracks, the total number of reviewers was 89.

3. Themes and Open Questions

The workshop discussions focused on the role, significance, current landscape, and research gaps related to metrics and evaluation in algorithmic fairness, especially in the context of advancements in generative AI and regulatory frameworks.

The panel discussion explored the complexities of defining fairness in generative AI, emphasizing the need for clear targets that account for the stakes of different domains, as fairness considerations vary between low- and high-stakes applications. While LLMs are often designed as general-purpose systems, fairness remains inherently domain-specific. The discussion highlighted the multidisciplinary nature of the field, stressing the importance of drawing insights from diverse areas of expertise. It was noted that policy and regulations should not be rushed but iteratively improved to keep pace with technological advancements. LLMs were also seen as an opportunity to build societally impactful applications. However, as previously observed with prediction and classification models, challenges arise here too when fairness objectives conflict with other performance metrics. Lastly, the panel addressed the question of whether AI redistributes harms and benefits, underscoring the need for methods to effectively capture and measure these shifts.

Below, we highlight some takeaways from the round table discussions and prospects for future work that we hope to address in future editions:

- *Metrics:* The roundtable on metrics examined the broader question of "Which metric when?" and highlighted key considerations. It emphasized the need for organizational-level policies and principles to guide metric selection, rather than leaving these decisions solely to individuals, who may lack the necessary expertise to choose appropriate metrics. The discussion also addressed the risks associated with turning metrics into targets, cautioning that this practice can lead to overfitting to specific metrics rather than effectively addressing underlying biases. For example, the use of leaderboards can sometimes encourage competitive optimization of metrics, which may undermine broader fairness goals.

- *Evaluation:* There are several important challenges in fairness evaluation. These include the complexity, scope, and scale of defining fairness and determining what and how to evaluate. A significant issue identified was the lack of inclusion and diversity in the design and development of evaluation frameworks, with limited involvement of experts from other fields and impacted stakeholders. The discussion highlighted a disconnect between research and real-life challenges, as research often fails to align with relevant regulatory frameworks or represent real-world scenarios. Additionally, there is a lack of appropriate structures and incentives to shift away from current practices and the fast pace of research, hindering investment in interdisciplinary and participatory approaches. Finally, the roundtable emphasized the limited openness in LLM research and underscored the need for stronger regulatory frameworks to mandate more rigorous audits and evaluations of models and datasets.

4. Acknowledgments

As AFME organizers, we extend our sincere thanks to all the invited speakers, panelists, and roundtable researchers, as well as the authors who contributed their work to the workshop. We also greatly appreciate the support of the 9 meta-reviewers and 89 reviewers, whose dedication and time were instrumental in maintaining the quality of the workshop content.

In alphabetical order, the meta-reviewers were: Ana-Andreea Stoica, Babak Salimi, Christoph Kern, Jessica Schrouff, Kun Zhang, Laurent Charlin, Mattia Cerrato, Stephen R Pfohl, Xueru Zhang.

The reviewers were: Abdelrahman Zayed, Adrián Arnaiz-Rodríguez, Afaf Taik, Agoritsa Polyzou, Alan Mishler, Aleksander Wieczorek, Aliasghar Khani, Aliasghar Khani, Amin Nikanjam, Andrés Domínguez Hernández, Anoush Najarian, Aparna Balagopalan, Arian Khorasani, Arian Khorasani, Canyu Chen, Chen Liang, Christine Herlihy, Daniela Cialfi, David Hartmann, David Kinney, Debashis Ghosh, Deborah D Kanubala, Dimitri Staufer, Eike Petersen, Elette Boyle, Elliot Creager, Esubalew Desta Asmare, Federico Peiretti, Georgia Baltsou, Gökhan Özbulak, Hadis Anahideh, Haolun Wu, Isabela Albuquerque, Isacco Beretta, Ishmeet Kaur, Jan Ramon, Jan Simson, Jiahao Li, Jonas Ngnawe, Julien Ferry, Kamorudeen A Amuda, Kate Donahue, Kimon Kieslich, Krystal Maughan, Mina Arzaghi, Maarten Buyl, Maarten Buyl, Marianne Abemgnigni Njifon, Marta Marchiori Manerba, Martin Lopatka, Martina Cinquini, MaryBeth Defrance, Matteo Fabbri, Matthew Landers, Mattia Cerrato, Megha Srivastava, Melissa Hall, Minyechil Alehegn tefera, Muhammad Mohsin, Otto Sahlgren, Peeyush Agarwal, Prakhar Ganesh, Prasanjit Dubey, Rajeev Ranjan Dwivedi, Rakshit Naidu, Ramya Srinivasan, Ranya Aloufi, Robin Burke, Saber Malekmohammadi, Samuel Dooley, Samuel R Mayworm, Sanghamitra Dutta, Sanne Vrijenhoek, Seamus Somerstep, Sebastian Zezulka, Shenao Yan, Shomik Jain, Sofia Jaime, Sri Sri Perangur, Stacey Truex, Sukanya Moorthy, Taofeek Abayomi, Tareen Dawood, Tareen Dawood, Tim Räz, Vidhya Kamakshi, Vishal Bhalla, Xuchen Li, Yanan Long, Zairah Mustahsan, Zeyu Tang, Zhiyu Guo, Ziqing Yang.

We would like to also thank the NeurIPS 2024 workshop chairs Bo Han, Manuel Rodriguez, Adil Salim, Rose Yu as well as the NeurIPS staff Terri Auricchio, Brad Brockmeyer, Lee Campbell, Tony Manzo, Brian Nettleton, Max Wiesner, and Stephanie Willes for their technical and organizational support.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 60–69, 2018.

A Arora, M Barrett, E Lee, Eivor Oborn, and K Prince. Risk and the future of ai: Algorithmic bias, data colonialism, and marginalization. *Information and Organization*, 33 (3):100478, 2023.

Mercy Nyamewaa Asiedu, Awa Dieng, Iskandar Haykel, Negar Rostamzadeh, Stephen Pfohl, Chirag Nagpal, Maria Nagawa, Abigail Oppong, Sanmi Koyejo, and Katherine Heller. The case for globalizing fairness: A mixed methods study on colonialism, ai, and health in africa. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2024.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, 2022.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 514–524, 2020.

Weixin Chen, Li Chen, Yongxin Ni, and Yuhan Zhao. Causality-inspired fair representation learning for multimodal recommendation, 2024.

Luca Deck, Jan-Laurin Müller, Conradin Braun, Domenique Zipperling, and Kühl Niklas. Implications of the AI act for non-discrimination law and algorithmic fairness. *European Workshop on Algorithmic Fairness*, 2024.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, and Aaron Roth. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76, 2021.

Zhoujie Ding, Ken Ziyu Liu, Pura Peetathawatchai, Berivan Isik, and Sanmi Koyejo. On fairness of low-rank adaptation of large models. *First Conference on Language Modeling*, 2024.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. *The Debates workshop at the 35th International Conference on Machine Learning*, 2018.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems, NeurIPS 2016*, pages 3315–3323, 2016.

Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 181–190, 2019.

Deborah Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2017.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1):3–32, 2024.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023.

Kristian Lum, Jacy Reese Anthis, Chirag Nagpal, and Alexander D'Amour. Bias in language models: Beyond trick tests and toward ruted evaluation. *CoRR*, abs/2402.12649, 2024.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. When causality meets fairness: A survey. *Journal of Logical and Algebraic Methods in Programming*, 141:101000, 2024.

Miriam Rateike. Algorithmic fairness over time: Advances & prospects. *European Workshop on Algorithmic Fairness*, 2024.

Matheus Schmitz, Rehan Ahmed, and Jimi Cao. Bias and fairness on multimodal emotion detection algorithms. *arXiv preprint arXiv:2205.08383*, 2022.

Lindsay Weinberg. Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ml fairness approaches. *The Journal of Artificial Intelligence Research*, 74, 2022.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333. PMLR, 2013.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 993–999, 2023.