# The Need for User-centred Assessment of AI Fairness and Correctness

Simone Stumpf
simone.stumpf@glasgow.ac.uk
University of Glasgow
UK

Evdoxia Taka
evdoxia.taka@glasgow.ac.uk
University of Glasgow
UK

Yuri Nakao
nakao.yuri@fujitsu.com
Fujitsu Ltd.
Japan

Lin Luo
l.luo.1@research.gla.ac.uk
University of Glasgow
UK

Ryosuke Sonoda
sonoda.ryosuke@fujitsu.com
Fujitsu Ltd.
Japan

Takuya Yokota
yokota-takuya@fujitsu.com
Fujitsu Ltd.
Japan

## ABSTRACT

AI needs to be fair and robust, especially to meet demands of new regulation. Regular assessments are key but it is unclear how we can involve stakeholders without a background in AI in these efforts. This position paper provides an overview of the problems in this area, discusses the current work and looks ahead to future research needed to make headway in user-centric assessment of AI.

## CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing systems and tools*; **Interactive systems and tools**; • **Computing methodologies** → **Learning settings**.

## KEYWORDS

AI assessment, human-in-the-loop AI, AI fairness

**ACM Reference Format:**
Simone Stumpf, Evdoxia Taka, Yuri Nakao, Lin Luo, Ryosuke Sonoda, and Takuya Yokota. 2024. The Need for User-centred Assessment of AI Fairness and Correctness. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '24), July 01–04, 2024, Cagliari, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3631700.3664912

## 1 THE PROBLEM AND OUR POSITION

AI can benefit people's work and everyday lives through decision support or recommender systems, curation of media content and AI-generated text or images. Yet, a significant barrier to reaping the benefits of AI is their unassessed potential for quality issues: lack of truthfulness, bias and robustness which may cause unfair outcomes, treatment and wider damage [12, 15, 16]. Hence, AI assessment has become necessary, in line with existing and impending regulatory frameworks (e.g., US President Biden's Executive Order, the UK AI Regulatory Bill, and the EU AI Act).



*UMAP Adjunct '24, July 01–04, 2024, Cagliari, Italy*

ACM ISBN 979-8-4007-0466-6/24/07
https://doi.org/10.1145/3631700.3664912

To date, a number of tools have been developed to help AI experts assess and mitigate bias and unfairness while maintaining accuracy [1, 2, 5, 16]. While great strides have been made in this direction, there are still a number of major issues to overcome:

- Bias, as an AI technical concept, does not necessarily lead to unfairness as perceived by users or in law. Fairness is often only evaluated for protected by (the UK) law characteristics (e.g. gender, age, marital status, race, disability etc.), but users are also sensitive to non-protected characteristics (e.g. income, financial stability etc.) [11] and other issues of equity and justice. Assessing bias and fairness often relies on technical measures related to ground truth "correctness". However, for many domains and applications, ground truth might not be fully aligned with users' perspectives of fairness [10].
- There is often a trade-off between accuracy and fairness, and even within fairness measures [23]. Users' notions of what is "fair" or "biased" differ markedly between domains and application contexts [9]. Even within user groups there is sometimes little consensus of how to define and achieve less biased, fairer outcomes [8].
- If we leave assessments to AI experts, they might not do this at all or choose measures that are technically easier to define and operationalize instead of fully capturing users' perspectives [3, 7]. This could lead to accusations of "fairwashing" when user involvement is lacking.
- However, giving users without AI expertise the power to assess AI is complex. Responsible AI-related concepts need to be made transparent and be communicated in an accessible way. Support needs to be given to users for assessment, and users need to be provided with ways to feedback on issues that need to be acted upon after they have assessed an AI system.

Thus, we argue and agree with others [4] that current ways of assessing AI are "broken", and we join in calls that aim to re-centre AI development – and assessment – around its users [19, 21]. We believe, alongside others [9, 14], that input from users is required to tackle these issues, however, approaches of seeking and integrating users' input are in their infancy and in urgent need of research.

## 2 CURRENT WORK

User input is required to develop responsible AI systems but there are many ways that this could be achieved. It has been suggested

**Table 1: Requirements arising from design workshops covered by conventional tools and ours, FairHIL [17]. In the Stakeholder row, L stands for the requirements from loan officers, D from data scientists, and B from both stakeholders. WiT, FV, FS, SV, FET, and HFIL stand for What-if Tool [25], FairVis [5], FairSight [1], Silva [26], Fairness Elicitation Tool [6], and FairHIL, respectively.**

| Area | Use | Requirement | Stake-holder | WiT | FV | FS | SV | FET | FHIL (ours) |
|---|---|---|---|---|---|---|---|---|---|
| 1. Attribute overviews | Informational | 1.1 attributes, number of records and attribute value distributions | B | ✓ | ✓ | ✓ | | ✓ | ✓ |
| | | 1.2 amount of missing data | B | | | | | | |
| | | 1.3 fairness metrics for model and individual protected attributes | B | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| | | 1.4 target distribution | B | ✓ | ✓ | | | | ✓ |
| | | 1.5 protected attributes | D | | | ✓ | ✓ | | ✓ |
| | | 1.6 explanation how attribute values are calculated or derived | D | | | | | | |
| | | 1.7 Identify if the data is subjective/ objective | D | | | | | | |
| | | 1.8 Identify where the data has come from (applicant, bank, third party) / attribute provenance | D | | | | | | |
| 2. Investigate Relationships between attributes | Informational | 2.1 distribution of protected attributes with other attributes | B | ✓ | | ✓ | | | ✓ |
| | | 2.2 distribution of user-selected attribute values (e.g. credit risk ratings) and target values | B | ✓ | | ✓ | | | ✓ |
| | | 2.3 distribution of two user-selected attributes' values | B | ✓ | | | | | ✓ |
| | | 2.4 Credit risk rating traffic light system | D | | | | | | |
| | Functional | 2.5 Support data transformations (e.g. categorical into numerical, binning) | D | ✓ | | | | | |
| | | 2.6 support filling in missing values | D | ✓ | | | | | |
| | | 2.7 Support creation of new attributes (i.e. calculated from other attributes e.g. affordability) | B | ✓ | | | | | ✓ |
| | | 2.8 Ability to create/include own fairness metric (if not already in system) | B | | | | | | ✓ |
| 2. Investigate Relationships between attributes | Functional | 2.9 Allow creation of subgroups based on a combination of attributes and see their distribution on target | B | ✓ | ✓ | | | ✓ | ✓ |
| | Adjust model | 2.10 Input custom thresholds to affect AI model | B | | | | | | |
| | | 2.11 Change weights on attributes to adjust AI model | B | ✓ | | | | | |
| | | 2.12 Remove attributes | B | ✓ | | | | | |
| | | 2.13 Change weightings of attribute variables (the variables that make up the attribute) on attributes | D | ✓ | | | | | |
| | | 2.14 Identify similar attributes which do not contain protected attributes and substitute attribute from these choices | D | | | | | | |
| | | 2.15 Optimise model against fairness metrics and accuracy automatically | D | | | | | | |
| | | 2.16 Feedback to data scientists on 'questionable' attributes that should not be used for decision-making | L | | | | | | |
| 3. Causal Graph | Informational | 3.1 Node weight and impact on target | B | | | | | | ✓ |
| | | 3.2 Relationships between nodes, their 'strength' and direction | B | | | | ✓ | | ✓ |
| | | 3.3 Explanation of how the graph was derived | B | | | | | | |
| | Adjust model | 3.4 The ability to remove nodes and relationships to adjust AI model | B | | | | | | |
| 4. individual cases | Informational | 4.1 specific application and attribute values | B | ✓ | | ✓ | ✓ | ✓ | ✓ |
| | | 4.2 fairness metric for individual case | B | | | | | | |
| | | 4.3 Level of similarity between cases | B | | | | | ✓ | ✓ |
| | | 4.4 Select specific cases to compare and show which attributes are similar | B | | | ✓ | | ✓ | ✓ |
| | | 4.5 Show decision boundaries | B | ✓ | | | | | ✓ |
| | Functional | 4.6 See 'What If' results on target based on changes to attribute values | L | ✓ | | | | | |
| 5. Model | Informational | 5.1 how model works | B | | | | | | ✓ |
| | | 5.2 how it was created, rationale for decisions in modelling | B | | | | | | |
| | | 5.3 who created it | B | | | | | | |

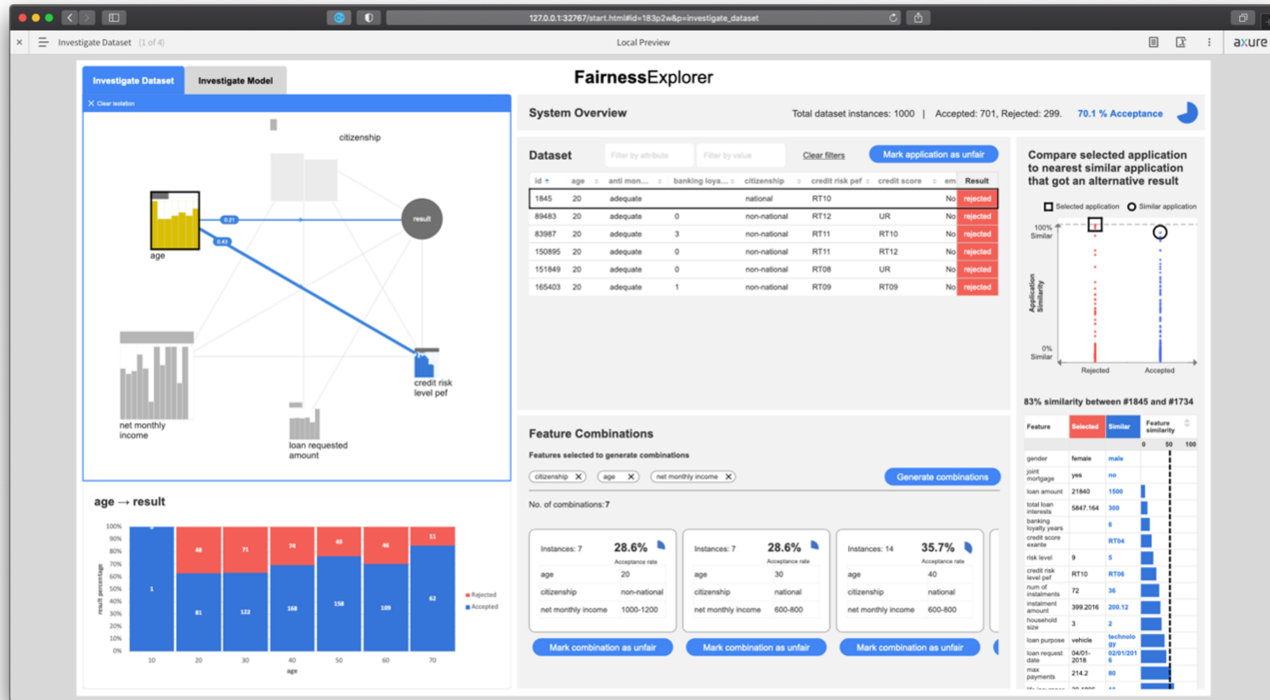


**Figure 1: FairHIL UI components (from [17])**

that users could be directly involved in deciding whether an AI model is fair: asking users to give feedback on preferred outcomes of decision scenarios instances and then "retrofitting" the appropriate fairness metric against the feedback [20]; showing a series of decision outcomes from two AI models against the ground truth and ask users to pick the preferred model to determine the fairness metric that fits the responses best [22, 27]; or allowing users to compare decision instances and their predicted outcomes, as well as providing information on model fairness employing different metrics [6].

Our own previous work has uncovered many ways in which to involve users without an AI background during the AI design and development lifecycle [24]. We argued that it is necessary to allow users' input from inception stages, allowing them to assess the business case and reasons for creating AI systems, through to collection and validation of training datasets, to assessment of models and their deployment.

To aid assessors without AI expertise, such as domain experts, we have investigated the design space of required information for them to make informed decisions about AI fairness [17]. This work suggested and evaluated User Interface (UI) components (Table 1) that should be available to these kinds of users during AI assessments. Fig. 1 shows UI components arising from this work to uncover biases and potential unfairness: causal graphs that explain how input and output features are related to each other, visualisations of outcomes for features and feature intersections, ways to show the weight of features in the AI model, as well as ways to compare instances to explore similarities and differences in predictions.

We have also explored whether it is feasible to support lay users in identifying "fair" and "unfair" decision instances and their outcomes [18], and using this feedback to improve fairness, inspired by our work in Explanatory Debugging to steer AI models [13]. We created a UI to support users in finding fairness issues and providing label and weight changes for features. Our research suggested that this kind of feedback can be used to increase the fairness of a retrained model for loan applications (by aggregating label and weight changes), measured through the Disparate Impact metric on the "Nationality" attribute. We also investigated cultural dimensions of fairness assessments in this work, and how this affected users' interactions and fairness assessments.

Currently, we are engaged in a series of studies on leveraging user feedback to make AI models fairer (currently in draft for publication). Our work seeks to investigate whether it is possible to identify what fairness metrics users choose and which attributes are involved, alongside how to integrate user feedback on fairness into personal and "merged" models. We have conducted two studies to collect user feedback and conducted a set of analyses to rigorously investigate the impact of user feedback using a large set of fairness metrics, allowing us to establish baselines, compare results, and delve deeply into the potential application and challenges of this approach. We have investigated, for example, using "fair" and "unfair" labels, as well as using feature weight changes obtained by users in retraining a model. We then observed the effect of this user input on group and individual fairness metrics. Our results show that we indeed see some fairness improvements on some metrics for some features but also that some individuals deteriorated fairness. We also found that other features seemed to matter but we have no firm grasp how users evaluate fairness on these features.

An additional strand of our work addresses how to support groups of users with negotiating and applying fairness to AI models (currently in draft for publication). To investigate this area, we designed an interactive system to explain various fairness metrics to individuals on protected attributes and for them to explore fairness. We then extracted their personal fairness metric preferences and used this as a basis for team negotiations. We found that perceptions of what is considered fair differs between stakeholders at the outset but that it is possible for them to achieve consensus on how fairness should be applied in the end.

## 3 FUTURE CHALLENGES AND DIRECTIONS

We are at the beginning of an exciting and unchartered time for AI, when more assessments are called for but the responsibilities will shift to users without specialist knowledge in AI. In participating in this workshop, we would like to share lessons learned from our research, engage in conversations around current and future research strands in user-centric AI, and network with other researchers and practitioners in this area. We anticipate discussing the following major research questions that warrant further reflection and work:

- How do end-users currently assess correctness and fairness, and the trade-offs, especially for generative AI solutions? What factors influence their fairness metric preferences?
- How can we reliably identify, prevent or counteract "gaming" fairness, i.e. a malicious actor abusing feedback making the AI less fair or feedback that violates regulation?
- How do we design and develop UIs, tools and associated methodologies to support end-users in assessing and mitigating AI, explaining responsible AI concepts?
- How can AI assessment be conducted in practice, especially to support collective fairness processes taking place in the wider user population?
- How can we translate results from "toy" research problems into the real world?
- How do we account for cultural diversity and shifts over time yet strive towards universal justice and equity for often marginalized groups?

## REFERENCES

[1] Yongsu Ahn and Yu-Ru Lin. 2019. FairSight: Visual Analytics for Fairness in Decision Making. *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1. https://doi.org/10.1109/TVCG.2019.2934262

[2] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (July 2019), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287 Conference Name: IBM Journal of Research and Development.

[3] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[4] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. http://arxiv.org/abs/2401.14462 arXiv:2401.14462 [cs].

[5] Á A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau. 2019. FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

[6] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu. 2021. Soliciting stakeholders' fairness notions in child maltreatment predictive systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[7] David De Cremer, Devesh Narayanan, Mahak Nagpal, Jack McGuire, and Shane Schweitzer. 2024. AI Fairness in Action: A Human-Computer Perspective on AI Fairness in Organizations and Society. *International Journal of Human–Computer Interaction* 40, 1 (2024), 1–3. https://doi.org/10.1080/10447318.2023.2273673 arXiv:https://doi.org/10.1080/10447318.2023.2273673

[8] Chowdhury Mohammad Rakin Haider, Christopher Clifton, and Ming Yin. 2024. Do Crowdsourced Fairness Preferences Correlate with Risk Perceptions?. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, USA) *(IUI '24)*. Association for Computing Machinery, New York, NY, USA, 304–324. https://doi.org/10.1145/3640543.3645209

[9] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3290605.3300830

[10] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. https://doi.org/10.1145/3531146.3533097

[11] Maria Kasinidou, Styliani Kleanthous, Pınar Barlas, and Jahna Otterbacher. 2021. I agree with the decision, but they didn't deserve this: Future Developers' Perception of Fairness in Algorithmic Decisions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 690–700. https://doi.org/10.1145/3442188.3445931

[12] Keith Kirkpatrick. 2016. Battling Algorithmic Bias: How Do We Ensure Algorithms Treat Us Fairly? *Commun. ACM* 59, 10 (Sept. 2016), 16–17. https://doi.org/10.1145/2983270

[13] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[14] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-centered approaches to fair and responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[15] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu,USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[17] Yuri Nakao, Lorenzo Strappelli, Simone Stumpf, Aisha Naseer, Daniele Regoli, and Giulia Del Gamba. 2022. Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. *International Journal of Human–Computer Interaction* (2022), 1–27. https://doi.org/10.1080/10447318.2022.2067936?src=

[18] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward Involving End-Users in Interactive Human-in-the-Loop AI Fairness. *ACM Trans. Interact. Intell. Syst.* 12, 3, Article 18 (jul 2022), 30 pages. https://doi.org/10.1145/3514258

[19] Yuri Nakao and Takuya Yokota. 2023. Stakeholder-in-the-Loop Fair Decisions: A Framework to Design Decision Support Systems in Public and Private Organizations. In *HCI in Business, Government and Organizations*, Fiona Nah and Keng Siau (Eds.). Springer Nature Switzerland, Cham, 34–46.

[20] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2020. How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations.

*Artificial Intelligence* 283 (2020), 103238. https://doi.org/10.1016/j.artint.2020.103238

[21] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504. https://doi.org/10.1080/10447318.2020.1741118 arXiv:https://doi.org/10.1080/10447318.2020.1741118

[22] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664

[23] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.

[24] Beatrice Vincenzi, Simone Stumpf, Alex S. Taylor, and Yuri Nakao. 2024. Lay User Involvement in Developing Human-Centric Responsible AI Systems: When and How? *ACM Journal on Responsible Computing* (March 2024). https://doi.org/10.1145/3652592 Just Accepted.

[25] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65. https://doi.org/10.1109/TVCG.2019.2934619

[26] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M. Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (, Honolulu, HI, USA,) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376447

[27] Takuya Yokota and Yuri Nakao. 2022. Toward a decision process of the best machine learning model for multi-stakeholders: a crowdsourcing survey method. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 245–254. https://doi.org/10.1145/3511047.3538033