Stakeholder Participation in AI Auditing: Challenges and Future Directions

Takuya Yokota
yokota-takuya@fujitsu.com
Fujitsu Limited
Kawasaki-city, Japan

Yuri Nakao
nakao.yuri@fujitsu.com
Fujitsu Limited
Kawasaki-city, Japan

Abstract

This position paper addresses the importance of involving diverse stakeholders in the development and auditing of AI systems to improve fairness and social acceptability. We review existing studies and identify challenges related to the needs of various stakeholders and the mitigation of biases arising from uneven or manipulated user feedback. Based on them, we highlight key challenges that require further research by mentioning ongoing work to enable stakeholders to provide feedback and audit AI models.

CCS Concepts

• **Human-centered computing** → *Collaborative interaction*; *User models*; *Computer supported cooperative work*.

Keywords

artificial intelligence, accountability, multi-stakeholder, fairness

ACM Reference Format:
Takuya Yokota and Yuri Nakao. 2025. Stakeholder Participation in AI Auditing: Challenges and Future Directions. In *Proceedings of Sociotechnical AI Governance workshop at CHI 2025 (STAIG @ CHI '25)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In the design and auditing processes of AI systems, involving a diverse range of stakeholders—such as developers, regulatory authorities, users, and civil society organizations—has become widely recognized as critical for integrating different values, reconciling interests, and ensuring social acceptability [5, 13, 21]. AI systems are deployed across various domains, including healthcare, finance, and facial recognition. On the other hand, issues of discrimination and unfairness caused by AI technologies are increasing. Because socially vulnerable groups are most affected by negative outcomes and are rarely included in AI decision-making [13], their unique needs and constraints often go unrecognized [5]. This omission can embed biases into the models, undermining efforts to address discrimination and unfairness. Consequently, users may lose trust in the system or refuse to adopt the technology altogether [21, 22]. In this position paper, we discuss challenges in developing frameworks and tools that facilitate the participation of diverse stakeholders in AI development and auditing with the goals of ensuring stakeholder representation, strengthening audit effectiveness, and enhancing the social acceptability of AI.

STAIG @ CHI '25, Yokohama, Japan

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

2 Needs for Multi-stakeholder Participation in AI Governance

In the context of AI governance, measures are required to incorporate the views of multiple stakeholders. In a study analyzing AI implementation in the public healthcare sector, Sun and Medaglia [21] found that three stakeholder groups, government policymakers, hospital administrators/physicians, and IT company managers, each identified distinct concerns. They concluded that the lack of a common problem awareness and conflicting interests impeded successful AI introduction [21]. Conflicts among the public's right to AI transparency, corporate intellectual property protections, and the privacy of data subjects have also been highlighted [11]. Excessive auditing may expose AI system vulnerabilities [3], while data disclosure poses privacy risks [24]. Keller et al. [11] documented cases in which the overuse of trade secrets by companies obstructed AI transparency, arguing that a framework that allows civil society organizations and independent oversight bodies to engage actively in discussions can protect audit objectivity and prevent hidden malpractice. Hence, successful AI adoption calls for governance mechanisms that promote collaboration among diverse stakeholders, facilitating the harmonization of different viewpoints.

Furthermore, studies highlight that the direct participation of diverse stakeholders in AI decision-making is beneficial. Lee et al. [13] established a computational model that incorporates stakeholder values and uses proxy voting, resulting in enhanced perceptions of fairness in decision-making and increased trust in the algorithm, suggesting improved social acceptance. Deng et al. [6] found that allowing users to engage in auditing increased opportunities to identify bias and errors in generative AI, enabled more timely detection of problems, and facilitated easier corrections based on user feedback. From these findings, involving diverse stakeholders in AI development and auditing processes can yield multiple benefits, including identifying problems from various perspectives, reconciling stakeholder interests, and strengthening social acceptance.

Several studies have proposed frameworks for involving diverse stakeholders in AI. For example, Human-in-the-loop (HITL) systems are designed to involve human feedback in AI model development, allowing iterative refinements based on user input [18]. These systems rely on visual interactive tools to help users recognize and mitigate biases in AI models by refining causal structures and addressing unfair causal relationships [7, 22]. However, building AI auditing frameworks involves addressing multiple issues.

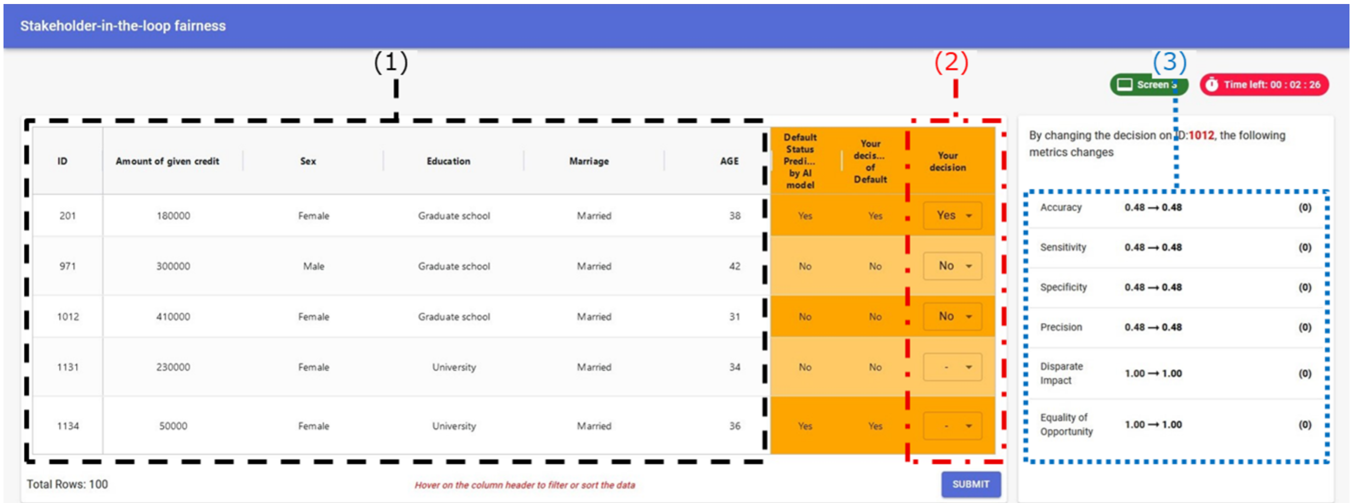

Figure 1: Interactive System Annotation Screen: This interface consists of three main components: (1) a dataset table for a credit card default scenario, (2) interactive buttons that allow users to annotate their decisions on the AI model's output, and (3) a performance and fairness metrics monitor that updates in real time based on these annotations.

A key requirement is providing each stakeholder group with the right level of information, tailored to their knowledge and concerns, so they can make informed decisions [14]. Individuals without a technical background often struggle to fully grasp AI systems' internal workings and may find it challenging to make informed judgments [1].

Thus, creating interpretable, interactive interfaces accessible to non-experts is essential [16]. The degree to which each stakeholder's opinion should be reflected is another key question. Studies on annotation-based AI auditing highlight that majority views can overshadow minority perspectives [13], and new methods are being explored to ensure more equitable representation [8]. Additionally, human feedback can introduce subjective biases [19] and may compromise fairness metrics. Taka et al. [22] demonstrated that annotation-based user feedback could worsen fairness scores if users focus on personal criteria, such as economic indicators, over accepted fairness norms. These findings underscore the difficulty of embedding fairness in AI while balancing individual input. Malicious actors may also exploit participatory approaches via data poisoning to manipulate outputs in ways that disproportionately affect certain groups [9]. Even a limited number of adversarial contributors can significantly degrade model accuracy in federated learning environments [23]. Integrity attacks, such as altering features or mislabeling data, can yield incorrect predictions and compromise auditing processes, including annotation-based feedback [10].

3 Challenges and Future Work

Building on this background, developing interactive tools that incorporate stakeholder feedback requires identifying the outcomes users expect from AI models and minimizing the risks of biased or strategically manipulated inputs. Previous research includes designing user interfaces that enable end users to evaluate the fairness of AI-based loan-screening models [16], interactive AI evaluation UIs aimed at reconciling fairness demands from both users and data scientists [15], and tools for stakeholders to provide direct model retraining feedback via annotations [22]. Additionally, a framework in which stakeholders provide feedback to facilitate AI model auditing and refinement [17] and a method for gathering user preferences regarding key metrics and, from multiple AI models, selecting the model that achieves the highest overall preference score [25], and the way of examination how different stakeholders hold preferences across multiple metrics and explored a preference-based method of defining stakeholder groups [26] have also been proposed.

Based on the research above, we developed a functioning interactive tool to facilitate the involvement of diverse stakeholders in AI auditing, as shown in Figure 1. The tool is designed to help identify the AI model that best satisfies the preferences of diverse stakeholders in real-world auditing scenarios. It accommodates various use cases involving multiple stakeholder roles. For example, in credit default prediction, the tool assumes participation from credit officers as decision-makers, credit card users as affected individuals, and financial auditors as regulators, each providing annotations from their respective perspectives. Users are guided to annotate a training dataset, from which the system infers latent stakeholder preferences as weights over multiple performance and fairness metrics. Using these inferred preferences, it estimates the most desirable AI model for each stakeholder group and selects the model that best matches the aggregated preferences of all participants. Ultimately, the tool aims to make stakeholder expectations visible in AI decision-making and to enable the selection of models that better reflect those expectations.

Despite these efforts, several open challenges and directions remain in balancing diverse stakeholder inputs. Many AI systems support a wide spectrum of stakeholders, from end users directly affected by AI decisions to data scientists and regulators. Each group brings different needs, expertise, and priorities, and existing frameworks often focus on user interfaces or preference elicitation but lack empirical evaluations of how these interactive methods translate into real-world outcomes. For example, increasing fairness for one demographic can reduce accuracy for another [4], underscoring the need to systematically measure the impacts of stakeholder feedback. Additionally, without transparent mechanisms for allocating the influence of feedback, minority viewpoints risk being overshadowed [8], and user trust can decrease. This imbalance can generate multiple adverse effects. First, those in the minority may perceive the process as unfair, eroding their trust in both the AI system and the organization deploying it. Second, the absence of explicit rules on how preferences are weighted can lead stakeholders to distrust the system, thereby reducing transparency [20]. In the student assignment case study, Robertson and Salehi [20] demonstrated that the system may fail to achieve adequate transparency and fairness, potentially further disadvantaging historically marginalized groups. Hence, it is crucial to understand how an AI system's final decisions, shaped by interactive tools, ultimately affect different stakeholders' interests. To enable such evaluations, we need methods that capture users' nuanced intentions and translate stakeholder preferences into quantifiable inputs.

Below are some key directions to address those issues:

- Quantify changes in AI performance across stakeholder groups.
- How do we systematically reconcile situations where improving outcomes for one group may inadvertently harm another?
- In what scenarios do certain experts' or vulnerable users' concerns warrant extra attention?
- Provide scenario-based evidence of how adjustments influence real-world decisions.

To address the key directions outlined above, we consider several possible extensions of our interactive tool, shown in Figure 1. First, we aim to extend the tool's model selection logic to better handle conflicts between stakeholder groups, for instance, by comparing different methods of aggregating preferences, such as majority voting, weighted scoring, or group-prioritized selection. Second, future studies could involve actual stakeholders using our interactive tool in real-world settings to assess how various model selection strategies influence their decisions and perceived fairness. Finally, we envision extending the tool to generative AI systems, such as large language models (LLMs) [2]. In this context, the PRISM dataset [12] offers insights into how different groups evaluate LLM outputs. The dataset includes detailed participant profiles (e.g., age, gender, religion, personal values) linked to evaluations of LLM responses across a wide range of conversation topics and model types. These data clarify when stakeholder preferences are in agreement and when they are in conflict, providing how we adapt our model selection strategies to ensure fair and acceptable outcomes for diverse users.

Overall, future research should aim to systematize how stakeholder perspectives are elicited, represented, and balanced in the development of AI systems. This includes integrating mechanisms for detecting bias and strategic manipulation, systematically evaluating real-world impact, and extending participation frameworks to emerging domains such as generative AI. Through such efforts, we can demonstrate that multi-stakeholder participation not only improves technical AI outcomes but also sustains the trust needed for widespread adoption.

4 Conclusion

This paper has identified challenges related to stakeholder participation in AI development and auditing, proposing future directions for frameworks and interactive tools that integrate diverse stakeholder values into AI model design. Existing frameworks and user interfaces for annotation-based feedback can demonstrate the value of enabling stakeholders to interact directly with models. Yet challenges remain, including how best to allocate influence among diverse viewpoints, safeguard against malicious manipulations, and measure real-world impacts on various demographic groups. By addressing these issues, we can ensure broader stakeholder representation in AI development, strengthen the effectiveness of auditing, and enhance the social acceptability of AI.

References

[1] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. doi:10.1145/3551624.3555290

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Sarah H. Cen and Rohan Alur. 2024. From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (San Luis Potosi, Mexico) *(EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, Article 13, 14 pages. doi:10.1145/3689904.3694711

[4] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. doi:10.1145/3097983.3098095

[5] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1571–1583. doi:10.1145/3531146.3533213

[6] Wesley Hanwen Deng, Wang Claire, Howard Ziyu Han, Jason I. Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. arXiv:2501.01397 [cs.HC] https://arxiv.org/abs/2501.01397

[7] Bhavya Ghai and Klaus Mueller. 2023. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 473–482. doi:10.1109/TVCG.2022.3209484

[8] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 115, 19 pages. doi:10.1145/3491102.3502004

[9] Isha Gupta, Hidde Lycklama, Emanuel Opel, Evan Rose, and Anwar Hithnawi. 2024. Fragile Giants: Understanding the Susceptibility of Models to Subpopulation Attacks. arXiv:2410.08872 [cs.LG] https://arxiv.org/abs/2410.08872

[10] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In *2018 IEEE Symposium on Security and Privacy (SP)*. 19–35. doi:10.1109/SP.2018.00057

[11] Perry Keller and Tanya F Aplin. 2024. Reconciling Trade Secrets and AI Public Transparency. (2024). doi:10.2139/ssrn.5044278

[12] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Dataset: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. arXiv:2404.16019 [cs.CL] https://arxiv.org/abs/2404.16019

[13] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 181 (Nov. 2019), 35 pages. doi:10.1145/3359283

[14] Yuri Nakao, Junichi Shigezumi, Hikaru Yokono, and Takuya Takagi. 2019. Requirements for Explainable Smart Systems in the Enterprises from Users and Society Based on FAT.

[15] Yuri Nakao, Lorenzo Strappelli, Simone Stumpf, Aisha Naseer, Daniele Regoli, and Giulia Del Gamba. 2023. Towards Responsible AI: A Design Space Exploration of Human-Centered Artificial Intelligence User Interfaces to Investigate Fairness. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1762–1788. doi:10.1080/10447318.2022.2067936 arXiv:https://doi.org/10.1080/10447318.2022.2067936

[16] 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 (July 2022), 30 pages. doi:10.1145/3514258

[17] 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.

[18] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (01 Mar 2018), 5–14. doi:10.1007/s10676-017-9430-8

[19] Charles Retzlaff, Srijita Das, Christabel Wayllace, Payam Mousavi, Mohammad Afshari, Tianpei Yang, Anna Saranti, Alessa Angerschmid, Matthew E. Taylor, and Andreas Holzinger. 2024. Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities. *J. Artif. Intell. Res.* 79 (2024), 359–415. https://api.semanticscholar.org/CorpusID:267448673

[20] Samantha Robertson and Niloufar Salehi. 2020. What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. *CoRR* abs/2007.06718 (2020). arXiv:2007.06718 https://arxiv.org/abs/2007.06718

[21] Tara Qian Sun and Rony Medaglia. 2019. Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly* 36, 2 (2019), 368–383. doi:10.1016/j.giq.2018.09.008

[22] Evdoxia Taka, Yuri Nakao, Ryosuke Sonoda, Takuya Yokota, Lin Luo, and Simone Stumpf. 2024. Human-in-the-loop Fairness: Integrating Stakeholder Feedback to Incorporate Fairness Perspectives in Responsible AI. arXiv:2312.08064 [cs.AI] https://arxiv.org/abs/2312.08064

[23] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. Data Poisoning Attacks Against Federated Learning Systems. arXiv:2007.08432 [cs.LG] https://arxiv.org/abs/2007.08432

[24] Shlomit Yanisky-Ravid and Sean K. Hallisey. 2019. "“Equality and Privacy by Design”: A New Model of Artificial Intelligen" by Shlomit Yanisky-Ravid & Sean K. Hallisey. 428-486 pages. https://ir.lawnet.fordham.edu/ulj/vol46/iss2/5

[25] 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. doi:10.1145/3511047.3538033

[26] Takuya Yokota and Yuri Nakao. 2023. Towards Multi-Stakeholder Evaluation of ML Models: A Crowdsourcing Study on Metric Preferences in Job-matching System. In *Computer-Human Interaction Research and Applications*, Hugo Plácido da Silva and Pietro Cipresso (Eds.). Springer Nature Switzerland.