Ensuring Artificial Intelligence is Safe and Trustworthy: The Need for Participatory Auditing

Patrizia Di Campli San Vito
Simone Stumpf
Patrizia.DiCampliSanVito@glasgow.ac.uk
Simone.Stumpf@glasgow.ac.uk
University of Glasgow
Glasgow, United Kingdom

Cari Hyde-Vaamonde
Gefion Thuermer
cari.1.hyde-vaamonde@kcl.ac.uk
gefion.thuermer@kcl.ac.uk
King's College London
London, United Kingdom

Abstract

Artificial intelligence (AI) is increasingly being used in many applications, yet governance approaches for these systems and applications are lagging behind. Recent regulations, such as the EU AI Act 2024, have highlighted the need for regular assessment of AI systems along their design and development lifecycle. In this context, *auditing* is critical to developing responsible AI systems, yet has typically been performed only by AI experts. In our work, we conduct fundamental research to design and develop auditing workbenches and methodologies for predictive and generative AI systems that are usable by stakeholders without an AI background, such as decision subjects, domain experts, or regulators. We describe our project to develop AI auditing workbenches and methodologies using co-design approaches, initial findings, as well as potential impacts of our work. We would like to share our experiences with the other workshop participants as well as discuss potential avenues for furthering the governance of AI systems.

CCS Concepts

• **Human-centered computing → Empirical studies in collaborative and social computing**; **User studies**; • **Computing methodologies → Artificial intelligence**.

Keywords

Predictive AI, Generative AI, Auditing, Co-Design, Harms

ACM Reference Format:
Patrizia Di Campli San Vito, Simone Stumpf, Cari Hyde-Vaamonde, and Gefion Thuermer. 2025. Ensuring Artificial Intelligence is Safe and Trustworthy: The Need for Participatory Auditing. In *Proceedings of CHI Conference on Human Factors in Computing 2025 Workshop 35 Sociotechnical AI Governance (CHI '25 Workshop).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction and Background

Artificial Intelligence (AI) is used extensively in people's work and everyday lives through supporting bail decisions [9], loan applications [16] or AI-generated text or images [17]. Yet, high-profile failures, such as predictive AI models that cast aside job applications by women or generative AI helping lawyers write judicial arguments that 'hallucinate' non-existing court cases, are grabbing the public's attention. A significant barrier to reaping the benefits of predictive and generative AI is their unassessed potential for harm. This has been echoed in regulatory frameworks, e.g. the EU AI Act 2024 [15], which call for responsible development processes to achieve safe and trustworthy AI through attention to accountability, transparency, and fairness. International frameworks such as the Hiroshima Policy Framework emphasize the responsibility of AI Actors to "promote safe, secure and trustworthy AI", and provide guidelines to ensure this happens. We focus on auditing as a "*systematic, independent and documented process for obtaining audit evidence and evaluating it objectively to determine the extent to which the audit criteria are fulfilled*" [11]. Though auditing is a long-established field, *AI auditing* is a much more recent topic that is attracting significant attention as regulations develop worldwide. Birhane et al. [4] conducted a systematic review of AI auditing and identified over 300 works over the past five years, suggesting a 4-stage process involving Harms Discovery, Standards Identification, Performance Analysis, and Audit Communication and Advocacy. *Harm discovery and impact assessments* for AI have attracted much research in recent years, partly driven by the emergence of the EU AI Act [5, 6]. For example, a number of risk and harm taxonomies have been developed [19, 22]. There are tools that support *performance analysis*, mainly by AI experts, for predictive decision-making [2, 21, 23] or generative AI [1]. Some tools have been developed to help non-experts to inspect fairness of AI models [7, 12, 13, 20, 24]. Yet, it has been suggested that AI auditing is "*a broken bus*" [4], without broader involvement by a range of stakeholders who directly use or are affected by AI system predictions, or those that regulate these systems [3, 10]. Our research intends to address this gap, by focusing on how to support *auditing* of AI systems by *non-experts*.

2 The Participatory Harm Auditing Workbenches and Methodologies (PHAWM) Project

Our project started in May 2024 and will last 47 months. It brings together a consortium of seven academic institutions — University

CHI '25 Workshop, April 27 2025, Yokohama, Japan

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

of Glasgow (lead), University of Edinburgh, King's College London, University of Sheffield, University of Stirling, Strathclyde University and University of York – and 25 partner organisations, such as Public Health Scotland, Fujitsu, Nokia Bell Labs, Scottish AI Alliance and many more. Our team consists of 20 academics, 14 post-doctoral researchers and 5 PhD students, from a variety of disciplines. We are focussing on the novel concept of participatory AI auditing [8] where a diverse set of stakeholders without a background in AI, such as domain experts, regulators, decision subjects and end-users, undertake audits of predictive and generative AI, either individually or collectively.

During the project we aim to build novel participatory auditing workbenches and methodologies for predictive and generative AI, targeted at diverse stakeholders. Through our research activities we will investigate and design novel interfaces to support participatory auditing, explore and develop new measures to assess AI that represent stakeholders' needs, and create approaches to involve a diverse set of stakeholders yet guard against harms by malicious actors. Through our work we hope to improve future AI development practices, affect the trajectory of new certification and regulatory frameworks for AI solutions as well as educate the public about the need for AI auditing and responsible AI.

Our work is carried out across two distinct streams: predictive AI and generative AI, in which we tackle two different use cases each. **Use Case 1: Health** investigates two models in the health sector, involving end users i.e. healthcare professionals, and decision subjects of the models. (i) Scottish Patients at Risk of Readmission and Admission (SPARRA) helps healthcare professionals by predicting a person's likelihood of being admitted to hospital as an emergency in-patient within the next year. (ii) The School Attachment Monitor (SAM) uses videos of children undergoing the Manchester Child Attachment Story Task to help with the identification of children with insecure attachment. **Use Case 2: Media Content** will explore two scenarios with moderators and users. (i) Search engine results (ii) Hate speech detection. **Use Case 3: Cultural Heritage** explores metadata generation and summarisations for collections of historical material with curators and users. **Use Case 4: Collaborative Content Generation** will investigate Wikipedia articles co-written by AI and editors, focusing on non-English languages and health.

These streams work in parallel, where Use Case 1 (predictive AI) and Use Case 3 (generative AI) have started already; Use Case 2 (predictive AI) and Use Case 4 (generative AI) will start later this year. Several institutions work together on each use case and partner institutions support the use cases in a plethora of ways, such as providing and discussing models or participating in meetings and on the advisory board. Currently, we are conducting user research to allow us to build a first prototype version of the workbench and methodology, which will then be evaluated through user studies. Here we focus on the results of current work to investigate auditing of SPARRA within Use Case 1 Health.

3 Initial Findings from Use Case 1 Health - SPARRA

We ran three co-design workshops focusing on SPARRA between 24th October and 18th December 2024, each lasting two hours and containing mostly group activities with some individual components to manage brainstorming contributions.

We followed state-of-the-art recommendations to recruit a diverse group of participants through posters, social media, and mailing lists of local community organisations. We selected 12 out of 96 prospective participants, aged between 20 and 74 years, maximising diversity and prioritising people from under-represented and marginalised groups. For Workshops 2 and 3, respectively 10 and 9 participants of Workshop 1 returned. Participants were compensated with £40 per workshop and the user research was approved by our institution's Ethics Committee.

Workshop 1 consisted of three sets of activities focusing on auditing a predictive AI system, SPARRA as a use case, and AI auditor examples. Participants were asked to identify when the system should be audited, what information they would need, and who should audit. We then asked participants to complete an impact assessment matrix associating who would be impacted by SPARRA scoring and the positive/negative nature of that impact. Facilitators collected the reported negative impacts as a list of potential harms, then groups discussed ideas on how these could be evaluated. Finally, participants proposed fictitious auditor personas [14]. ***Workshop 2*** focused on harm prioritisation and measure development. Facilitators introduced a consolidated list of seven harms identified in Workshop 1 and participants rated each harm as to likelihood and impact, prioritising them. Facilitators then introduced measures and their use as audit criteria, as well as the measures in use for the performance and fairness assessment of SPARRA. Participants were then asked to develop measures for the prioritised harms, first individually and then in group discussions. ***Workshop 3*** explored user journeys and user interfaces (UI). Facilitators used low-fidelity prototyping techniques [18] to discuss required screens, content, UI layout and components in groups.

Participants stressed the importance of continuous audits, especially at every major change of the planned system and before release, to improve quality, ethics, privacy and security of the AI system. They specified 17 personas as archetypes of auditors and named a remarkably diverse range of stakeholders to be involved at each stage of the AI development lifecycle, including the general public/end-users, government agencies/regulators and process owners. The majority of identified stakeholders were external auditors, and participants stated that they needed access to a very extensive and diverse set of information on the system to carry out an audit, including policies, e.g. privacy, sustainability, transparency or deployment & IP fair use description, and detailed information about data and performance.

Our results showed that participants were able to articulate harms relatively easily but perhaps not very precisely and concisely. When participants prioritised their list of harms, all harms were clustered together fairly tightly, but Harm 1 ("*Inaccurate scores for patients with low resource access can lead to untimely diagnoses*") was clearly judged to be the most important. We therefore focused on measuring this harm, and several metrics for it were proposed. As variables in these metrics, participants wanted to use income or employment status information as a proxy for the resource access level of a patient; note, however, that this information is not used as part of the SPARRA model nor tracked as part of the underlying data. Participants looked mainly to relatively simple metrics to apply, such as parity, for example, parity between average SPARRA scores for patients with low resource access and for patients whose resource access is not low, with the assumption that patients with low resource access provide insufficient data in the SPARRA system leading to inaccurate SPARRA scores. We found that it was often not easy for them to formulate concrete measures which were directly related to the harms and developing metrics took participants considerable time.

When we investigated how participants imagined stepping through the system as a user journey, we found that the main flow was similar between participant groups, with minor changes reversing the order of some screens or introducing additional screens to help the auditing process. Participants indicated that a pre-existing taxonomy of harms would be helpful. For harms developed and applied to the system, participants wanted to carry out prioritisation as to the harm's impact and likelihood, similar to what they did in Workshop 2. Our experience from our workshops suggests that stakeholders need additional support to identify and develop harms, as well as setting up metrics for the performance analysis. To set up metrics, participants requested examples and detailed definitions of the metrics. Once the system is tested against the metrics, i.e. performance analysis of the system is carried out, participants told us that they wanted to know the progress of an audit, more information on the data used for measuring the performance on a metric, and give an indication of what is considered effective performance, i.e., a threshold setting which passes the audit. The results of the co-designed UIs indicated the importance of clear navigation and menu structure. In addition, harms and metrics UIs need to be carefully designed, to allow harms to be specified easily and metrics to be associated with harms. Information needs to be displayed on specific metrics, which should include a simple visual presentation of the metric outcomes, but there also needs to be an overview of all metrics.

4 Discussion and Conclusion

Our results have four implications for auditing and the design of auditing tools that we would like to discuss at the workshop. First, how can we put auditors without AI knowledge in charge? Many of our participants found it hard to articulate harms and metrics useful in an audit and thus there needs to be considerable support to guide them through the auditing methodology and through the auditing tools. Second, while our participants identified a range of interesting potential harms of SPARRA, they also indicated that they were interested in existing taxonomies and metrics. We suggest both are needed but there needs to be more discussion about which taxonomies are applicable and how they should be used within the auditing tools. Third, it also became clear that some harms could not be measured by using information currently collected or shared. Thus, this needs to be integrated into the auditing feedback, and might require AI system developers to collect further data to progress the audit and assess the system beyond simple performance of the model predictions. Last, there is a fine balance between providing the information needed to conduct an audit without overwhelming auditors without domain or technical knowledge. Careful UI design is thus necessary, translating the principles and requirements established in the workshops into requirements and wireframes.

Leading on from these results, we will seek the input of other stakeholders who might audit SPARRA, such as General Practitioners (GPs), and we will further analyse the harms identified by stakeholders and their relationships to harm taxonomies. Further investigation of regulations and policies, including industry codes of conduct, will provide useful guidance for how AI developers can embed such frameworks in their development practice, and inform the practical requirements around implementing PHAWM's participatory auditing processes. Our next steps will involve the detailed design, implementation and evaluation of auditing tools that will empower stakeholders with diverse backgrounds, yet without an AI background, to audit AI systems successfully. Our work provides important lessons for ensuring the responsible design of AI systems.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0011).

References

[1] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 304, 18 pages. https://doi.org/10.1145/3613904.3642016

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

[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)*. 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. *Proceedings - IEEE Conference on Safe and Trustworthy Machine Learning, SaTML 2024* (2024), 612–643. https://doi.org/10.1109/SaTML59370.2024.00037

[5] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. AI Design: A Responsible Artificial Intelligence Framework for Prefilling Impact Assessment Reports. *IEEE Internet Computing* 28, 5 (Sept. 2024), 37–45. https://doi.org/10.1109/MIC.2024.3451351 Conference Name: IEEE Internet Computing.

[6] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. https://doi.org/10.48550/arXiv.2306.03280 arXiv:2306.03280 [cs].

[7] 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* (Yokohama, Japan) *(CHI '21)*. https://doi.org/10.1145/3411764.3445308

[8] Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–19. https://doi.org/10.1145/3491102.3517441

[9] Christoph Engel, Lorenz Linhardt, and Marcel Schubert. 2024. Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law* (Feb. 2024). https://doi.org/10.1007/s10506-024-09389-8

[10] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, 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*. 1–16.

[11] ISO 19011:2018 1998. *Guidelines for auditing management systems* (3 ed.). Standard. International Organization for Standardization, Geneva, CH.

[12] 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. https://doi.org/10.1080/10447318.2022.2067936 arXiv:https://doi.org/10.1080/10447318.2022.2067936

[13] 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

[14] Timothy Neate, Aikaterini Bourazeri, Abi Roper, Simone Stumpf, and Stephanie Wilson. 2019. Co-Created Personas: Engaging and Empowering Users with Diverse Needs Within the Design Process. , 12 pages. https://doi.org/10.1145/3290605.3300880

[15] European Parliament. 2023. *EU AI Act: first regulation on artificial intelligence*. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (accessed 09/12/2024).

[16] Erasmo Purificato, Flavio Lorenzo, Francesca Fallucchi, and Ernesto William De Luca. 2023. The Use of Responsible Artificial Intelligence Techniques in the Context of Loan Approval Processes. *International Journal of Human–Computer Interaction* 39, 7 (April 2023), 1543–1562. https://doi.org/10.1080/10447318.2022.2081284

[17] Janet Rafner, Blanka Zana, Peter Dalsgaard, Michael Mose Biskjaer, and Jacob Sherson. 2023. Picture This: AI-Assisted Image Generation as a Resource for Problem Construction in Creative Problem-Solving. In *Creativity and Cognition*. ACM, Virtual Event USA, 262–268. https://doi.org/10.1145/3591196.3596823

[18] Atiqur Rahaman. 2024. *40 User Interface Elements: Must to Learn For Every Designer*. https://designmonks.co/user-interface-elements/ (accessed 16/12/2024).

[19] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. arXiv:2408.12622 [cs.AI] https://arxiv.org/abs/2408.12622

[20] Qianwen Wang, Zhenhua Xu, Chen Zhu-Tian, Yong Wang, Shixia Liu, and Huamin Qu. 2021. Visual Analysis of Discrimination in Machine Learning. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1470–1480. https://doi.org/10.1109/TVCG.2020.3030471

[21] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. arXiv:2303.16626 [cs.LG]

[22] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986 [cs.AI] https://arxiv.org/abs/2310.11986

[23] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[24] Angie Zhang, Olympia Walker, Kaci Nguyen, Jiajun Dai, Anqing Chen, and Min Kyung Lee. 2023. Deliberating with AI: Improving Decision-Making for the Future through Participatory AI Design and Stakeholder Deliberation. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 125 (apr 2023), 32 pages. https://doi.org/10.1145/3579601