Creating an Effective Methodology for End-User Engagement in AI Auditing

Emily O'Hara[1][0000−0002−9798−9307], Eva Fringi[2][0009−0008−9642−660X], and Kathryn Simpson[1][0000−0002−6883−4146]

[1] Digital Humanities Institute, University of Sheffield, Sheffield, UK
{e.w.ohara,kathryn.simpson}@sheffield.ac.uk
[2] Glasgow Interactive SysTems (GIST), Computing Science, University of Glasgow, Glasgow, UK
Evangelia.Fringi@glasgow.ac.uk

Abstract. A methodology explains the object of an AI-audit. This object has three loci: identifying significant events (harms or risks), governance (model is behaving as expected), and assurance (trust). The methodology in this paper is being developed as part of the PHAWM project,[3] which seeks to design a workbench that supports inclusive, participant-led auditing of AI application across a range of domains. Project participants range from health service users, parents of school-aged children, to museum professionals and librarians. The project addresses a key gap in existing approaches: the absence of human-centred infrastructures that empower end-users to identify events,[4] understand system behavior and participate meaningfully in audit processes.

Keywords: Participatory Design · Explainable AI (XAI) · Methodology · Trust and Ethics in AI systems · Audit and Accountability

1 Introduction

As artificial intelligence (AI) systems become embedded in decision-making across sectors, the need to audit these systems for fairness and accountability grows increasingly urgent. In this paper we present a prototype dynamic methodology for AI auditing. Our focus has been on creating an understandable and standardised methodology that guides auditors and audit instigators regardless of AI development expertise to think through and understand the auditing process. In doing so, we contribute a user-centred, literature-informed framework tailored for practical application in AI auditing.

[3] The Participatory Harm Auditing Workbenches and Methodologies project can be found at https://phawm.org

[4] An event refers to an occurrence triggered by an AI application that may affect entities and has associated metrics. Each event can be assessed for likelihood, magnitude, and positive or negative valence. We avoid the term harm in our methodology due to its subjectivity, although we acknowledge its common use, including in our own project title, within AI auditing discourse.

2 Related Work

The methodology of an AI audit workbench has been defined as an organising process to enable the audited AI resource's "past or present behaviour" to be understood in a structured manner [14]. But, researchers have pointed out that audits can exacerbate harm if there is no consensual understanding of what an audit is, what is it for, how it should be enacted, and what standards are used to support it [14]. It has been identified that even within the technical and development spaces of AI there is a lack of AI auditing competence [12]. Further to this, as [13] shows, the processes, rules, structures, languages, design and implementation of an audit methodology can reify hegemonic digital power structures by being presented as absolute. We also know that the methodological characteristics of an audit are what define it, and as such they are frequently driven by regulatory pressure or by private companies' needs to manage their perceived risks and are therefore non-transferable or replicable [14]. Thus we argue that regardless of competence level, there is a need for a clear, explainable, and repeatable methodology framework.

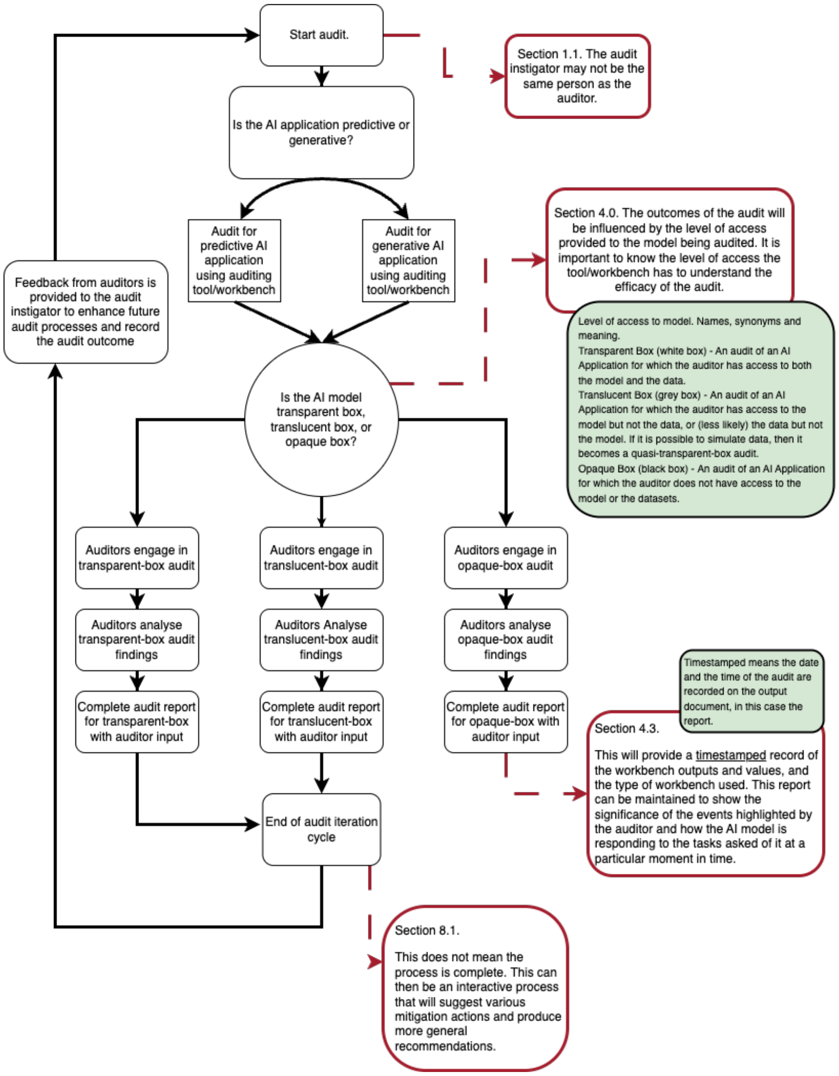
According to [8], harm can occur because "...existing structures around algorithmic systems often fail to empower those who directly interact with and are affected by AI resources or tools". At the same time, case studies have shown that end-users can identify harms which formal testing processes have missed [1,4,6,7,10,11,18]. As [7] evidences, even though expert-led audits have impact, they encounter major blindspots in the absence of everyday use context. They go on to argue for greater study in this area and propose a human-centric audit structure predicated on their study of harm identification and understanding by end-users.

[16] contend that accountability is the primary goal of AI auditing, but that little auditing infrastructure actually supports it. They explain that accountability is essential for harm identification and prevention, as it allows pinpointing causes and affected parties, however they observe that tools which help reveal harm or communicate the audit result are lacking across current AI audit provisions. Their findings also indicate that audit creators have difficulty engaging affected stakeholders. In this instance, [16] argue for the importance of "auditor independence, data access, peer review, standardisation, and advocacy" as ways to embed responsibility and accountability in auditing systems. They note that the development of participatory methods for auditing is a promising new direction in accountability and urge policymakers to include participation as a requirement in audit guidance. Our proposed methodology provides the infrastructure and mechanisms necessary to support meaningful end-user participation in auditing.

3 Framework Design and Methodological Approach

3.1 Diagram Structure

We present a dynamic diagrammatic methodology for AI auditing, developed through the PHAWM project, with the express functions of being understand-


Start audit.
Is the AI application predictive or generative?
Audit for predictive AI application using auditing tool/workbench
Audit for generative AI application using auditing tool/workbench
Is the AI model transparent box, translucent box, or opaque box?
Auditors engage in transparent-box audit
Auditors engage in translucent-box audit
Auditors engage in opaque-box audit
Auditors analyse transparent-box audit findings
Auditors Analyse translucent-box audit findings
Auditors analyse opaque-box audit findings
Complete audit report for transparent-box with auditor input
Complete audit report for translucent-box with auditor input
Complete audit report for opaque-box with auditor input
End of audit iteration cycle
Feedback from auditors is provided to the audit instigator to enhance future audit processes and record the audit outcome

Section 1.1. The audit instigator may not be the same person as the auditor.

Section 4.0. The outcomes of the audit will be influenced by the level of access provided to the model being audited. It is important to know the level of access the tool/workbench has to understand the efficacy of the audit.

Level of access to model. Names, synonyms and meaning.
Transparent Box (white box) - An audit of an AI Application for which the auditor has access to both the model and the data.
Translucent Box (grey box) - An audit of an AI Application for which the auditor has access to the model but not the data, or (less likely) the data but not the model. If it is possible to simulate data, then it becomes a quasi-transparent-box audit.
Opaque Box (black box) - An audit of an AI Application for which the auditor does not have access to the model or the datasets.

Timestamped means the date and the time of the audit are recorded on the output document, in this case the report.

Section 4.3.

This will provide a timestamped record of the workbench outputs and values, and the type of workbench used. This report can be maintained to show the significance of the events highlighted by the auditor and how the AI model is responding to the tasks asked of it at a particular moment in time.

Section 8.1.

This does not mean the process is complete. This can then be an interactive process that will suggest various mitigation actions and produce more general recommendations.

Fig. 1. A methodological framework for AI auditing. The diagram was initially conceived by Cari Hyde-Vaamonde as part of PHAWM, and enhanced by drawing from other methodological diagrams, such as [20]. Red boxes appear on click and green on mouseover; shown here for context.

able to end-users and applicable in multiple audit scenarios. Unlike conventional text-dense audit procedures, this framework takes the form of a visual, decision-based diagram (see Fig. 1). It illustrates key decision nodes and branching pathways (in black), clickable pop-out references to supporting documentation (in red) and hoverable tooltips with definitions (in green), which for illustrative purposes are all actively presented. Designed to guide auditors, particularly non-experts, it is rooted in infrastructural participatory design principles like transparency [5] and explainability [16], in order to support critical decision-making by balancing navigability and simplicity with methodological rigour.

The diagram is structured around two primary branches: (1) whether the AI application is predictive or generative, and (2) the level of system access available: transparent, translucent, or opaque. These distinctions, drawn from established audit typologies [13,2,15], accommodate diverse audit contexts and access levels. The first branch acknowledges the differences inherent to each type of AI application. Predictive systems (e.g., risk scoring) raise concerns like bias and fairness, while generative systems (e.g., content creation) raise events such as misinformation or offensive outputs. Given these systems' differing logic and societal impacts, they require distinct audit strategies and evaluative criteria. The second branch, system transparency, guides users through audit tactics based on access constraints. This approach negates some of the perceived issues with a one-size-fits-all methodology and supports context-sensitive auditing.

Each diagram pathway outlines actionable steps (e.g., identify event, analyse outputs, complete reports). This is supported by interactive references offering procedural guidance, linking users to more in-depth documentation where needed, as well as tooltips containing definitions of technical language. In this way, the diagram aligns with calls for standardised language [9], and the embedded guidance reflects findings from [7,18] on the importance of platform affordances for user-led harm identification. Together, the static and dynamic elements serve distinct but complementary functions: the diagram shows the audit by mapping its structure, while the interactive features explain it by providing contextual detail.

3.2 Discussion of Approach and Implementation

Audit literature calls for effective participatory frameworks, yet many remain too complex for everyday users, with expert control limiting broader adoption. The PHAWM methodology counters this with an adaptive, iterative design that enables user-driven processes. As such, it is suitable for use within or outside formal audit regimes.

The diagram was designed to highlight key decision points (e.g., AI application type, model transparency) in an accessible branching format. It was refined through stakeholder workshops which focused on different AI applications and gathered user stories. The participants involved were recruited from various areas within the scope of intended users[5]. By embedding user agency, this process is designed to avoid "participation-washing" and respond to real-world audit needs[3].

Though visually simple, the diagram is underpinned by a detailed methodology accessible via tooltips and linked resources. Expert users can engage at

[5] For example healthcare service users for a healthcare app, parents and educators for a child psychology app, librarians and cultural heritage (CH) professionals for a meta data enrichment model.

depth, while others benefit from an interpretable and intuitive interface. By favouring a visual over text-based structure, the diagram lowers entry barriers and addresses procedural opacity. As [19] argue, effective end-to-end auditing requires shared, interactive infrastructures; similarly, [17] stress frameworks must be usable by interdisciplinary teams and auditable themselves. The PHAWM diagram addresses both, functioning as a practical tool and standardised schema that can be printed, embedded in software, or used in workshops.

4 Conclusions and Future work

This paper has introduced a participatory, event-focused auditing methodology in the form of a visual, decision-based diagram. While the framework is still in its early stages, its primary contribution lies in foregrounding explainability and accessibility as essential components of AI auditing, particularly for including and supporting non-expert auditors. Rather than presenting a fully realised solution, the diagram demonstrates how audit processes might be made more navigable and interpretable through visual and interactive design without compromising rigour. Future work will involve testing across diverse domains, refining the methodology based on user feedback, and integration with an online workbench to support collaborative end-user participation in auditing.

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). We would also like to thank our PHAWM colleagues.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Attenberg, J., Ipeirotis, P., Provost, F.: Beat the machine: Challenging humans to find a predictive model's “unknown unknowns”. Journal of Data and Information Quality (JDIQ) **6**(1), 1–17 (2015)
2. Benbouzid, D., Plociennik, C., Lucaj, L., Maftei, M., Merget, I., Burchardt, A., Hauer, M.P., Naceri, A., van der Smagt, P.: Pragmatic auditing: A pilot-driven approach for auditing machine learning systems (2024)
3. Birhane, A., Steed, R., Ojewale, V., Vecchione, B., Raji, I.D.: AI auditing: The Broken Bus on the Road to AI Accountability (Jan 2024)
4. Cabrera, Á.A., Druck, A.J., Hong, J.I., Perer, A.: Discovering and validating ai errors with crowdsourced failure reports. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2), 1–22 (2021)
5. Cen, S.H., Alur, R.: From transparency to accountability and back: A discussion of access and evidence in ai auditing (2024)
6. Deng, W.H., Guo, B., Devrio, A., Shen, H., Eslami, M., Holstein, K.: Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–18. ACM, Hamburg Germany (Apr 2023)

7. DeVos, A., Dhabalia, A., Shen, H., Holstein, K., Eslami, M.: Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In: CHI Conference on Human Factors in Computing Systems. pp. 1–19. ACM, New Orleans LA USA (Apr 2022)
8. DeVrio, A., Eslami, M., Holstein, K.: Building, Shifting, & Employing Power: A Taxonomy of Responses From Below to Algorithmic Harm. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 1093–1106. ACM, Rio de Janeiro Brazil (Jun 2024)
9. Kingsley, S., Zhi, J., Deng, W.H., Lee, J., Zhang, S., Eslami, M., Holstein, K., Hong, J.I., Li, T., Shen, H.: Investigating What Factors Influence Users' Rating of Harmful Algorithmic Bias and Discrimination. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing **12**, 75–85 (Oct 2024)
10. Lam, M.S., Gordon, M.L., Metaxa, D., Hancock, J.T., Landay, J.A., Bernstein, M.S.: End-user audits: A system empowering communities to lead large-scale investigations of harmful algorithmic behavior. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–34 (2022)
11. Li, R., Kingsley, S., Fan, C., Sinha, P., Wai, N., Lee, J., Shen, H., Eslami, M., Hong, J.: Participation and Division of Labor in User-Driven Algorithm Audits: How Do Everyday Users Work together to Surface Algorithmic Harms? In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–19. ACM, Hamburg Germany (Apr 2023)
12. Li, Y., Goel, S.: Artificial intelligence auditability and auditor readiness for auditing artificial intelligence systems. International Journal of Accounting Information Systems **56**, 100739 (2025)
13. Muldoon, J., Wu, B.A.: Artificial Intelligence in the Colonial Matrix of Power. Philosophy & Technology **36**(4), 80 (Dec 2023)
14. Mökander, J.: Auditing of AI: Legal, Ethical and Technical Approaches. Digital Society **2**(3), 49 (Dec 2023)
15. Mökander, J., Schuett, J., Kirk, H.R., Floridi, L.: Auditing large language models: a three-layered approach. AI and Ethics **4**(4), 1085–1115 (Nov 2024)
16. Ojewale, V., Steed, R., Vecchione, B., Birhane, A., Raji, I.D.: Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–29 (2025)
17. Raji, I.D., Buolamwini, J.: Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Communications of the ACM **66**(1), 101–108 (Jan 2023)
18. Shen, H., DeVos, A., Eslami, M., Holstein, K.: Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. Proceedings of the ACM on Human-Computer Interaction **5**(CSCW2), 1–29 (Oct 2021)
19. Waltersdorfer, L., Ekaputra, F.J., Miksa, T., Sabou, M.: AuditMAI: Towards An Infrastructure for Continuous AI Auditing (Jun 2024)
20. Waltersdorfer, L., Sabou, M.: Leveraging knowledge graphs for ai system auditing and transparency. Journal of Web Semantics (Dec 2025)