Comment on FR Doc # 2023-28232
Mozilla Open Source Audit Tooling (OAT) Project

Inioluwa Deborah Raji, Abeba Birhane, Briana Vecchione,
Ryan Steed, and Victor Ojewale

The rise of generative AI products has introduced a whole new set of difficulties in the landscape of AI evaluation, assurance, and accountability. Without clear reporting of appropriate application scope[1] and with little restriction on use cases, it can be difficult to more precisely identify and measure the prevalence of major issues. However, concrete examples of potentially harmful failures are already visible[2]: large language models making critical mistranslations in medical settings[3] and at the border[4]; text-to-image models amplifying representations of problematic cultural stereotypes,[5] and AI-powered image editing tools modifying images to become more sexually explicit or inappropriately racialized.[6] These examples represent only a few of many risks associated with generative AI, from discrimination to misinformation to negative environmental impacts.[7]

---

[1] Bommasani, Rishi, et al. "The foundation model transparency index." *arXiv preprint arXiv:2310.12941* (2023).
[2] Raji, Inioluwa Deborah, et al. "The fallacy of AI functionality." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.
[3] Mehandru, Nikita, Samantha Robertson, and Niloufar Salehi. "Reliable and safe use of machine translation in medical settings." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.
[4] Bhuiyan, J. "Lost in AI translation: Growing reliance on language apps jeopardizes some asylum applications." *The Guardian* (2023).
[5] Sasha Luccioni et al., "Stable Bias: Evaluating Societal Representations in Diffusion Models," 2023, https://openreview.net/forum?id=qVXYU3F017; Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023; Steed, Ryan, and Aylin Caliskan. "Image representations learned with unsupervised pre-training contain human-like biases," *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021; https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/; https://www.npr.org/sections/goatsandsoda/2023/10/06/1201840678/ai-was-asked-to-create-images-of-black-african-docs-treating-white-kids-howd-it-
[6] Heikkilä, Melissa. "The viral AI avatar app Lensa undressed me—without my consent." *Technology Review. https://www. technologyreview. com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent* (2022).; https://www.businessinsider.com/student-uses-playrgound-ai-for-professional-headshot-turned-white-2023-8; https://arstechnica.com/information-technology/2022/12/lensa-ai-app-causes-a-stir-with-sexy-magic-avatar-images-no-one-wanted/.
[7] Weidinger, Laura, et al. "Ethical and social risks of harm from language models." *arXiv preprint arXiv:2112.04359* (2021).

Most notably, the need to independently assess these systems via ***audits*** has been raised as a key direction in identifying and regulating concerns related to AI's widespread availability and use, and NIST's contribution to guidance and benchmarks for AI auditing (RFI #1.a) and standards for AI systems (RFI #3) will set a precedent for these efforts.

As part of the Mozilla Open Source Audit Tooling (OAT) project, we analyzed nearly 400 tools and resources being used by those in the algorithmic auditing space, interviewed over 20 algorithm audit practitioners to identify pain points in their practice, and analyzed over 300 case studies of algorithm audit investigations. This project is led by Mozilla Fellow Inioluwa Deborah Raji and completed in collaboration with a team of interdisciplinary algorithm audit scholars, Briana Vecchione, Abeba Birhane, Ryan Steed, and Victor Ojewale.

Several of our findings, including our taxonomies of AI auditing and AI audit tooling, may be useful to NIST as it develops guidelines and standards for generative AI and other kinds of AI systems.

1) What is an audit? Supporting both internal and external auditing.

Meaningful sociotechnical-informed evaluations[8] are necessary to be able to make progress in the identification and mitigation of harms arising from AI systems. At a minimum, these assessments, when done reliably, can inform the definition of contexts of premature use. Ideally, these evaluations are also ***independent*** of the engineering team, tied to a defined ***audit target*** and ***articulated expectations***. Most importantly, its outcome should be ***motivated by the objective of accountability***.[9] In other words: an ***audit***.

Crucially, audit work is not only conducted by internal teams or contracted consultants. In our work, we use the taxonomy of "internal" and "external" auditors to distinguish the two relatively distinct populations of practitioners investigating algorithmic harms.

Internal auditors are those who conduct an independent review of the development and deployment of the products in use, often through a contractual relationship with the audit target. Internal auditors typically operate under professional obligation and their

[8] Weidinger, Laura, et al. "Sociotechnical safety evaluation of generative ai systems." *arXiv preprint arXiv:2310.11986* (2023).
[9] Birhane, Abeba, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. "AI Auditing: The Broken Bus on the Road to AI Accountability." Forthcoming in the IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 2024, https://arxiv.org/abs/2401.14462

processes are designed to validate procedural expectations, minimize liability, and test for compliance to corporate principles and legal constraints.[10]

External auditors are not contractually linked with the audit target, and typically engage in investigations on the behalf of represented constituents' interests. External audit processes tend to be voluntary and aim for a material change in the situation (i.e., product updates, policy changes, recalls, etc.) to minimize the harm being experienced by those they represent.[11]

Recommendations for NIST

1. **When discussing and defining AI "audit" methods and standards (RFI #1.a.2, #3), maintain a focus on independence, specific audit targets, and articulated expectations.** These elements are crucial components of accountability.
2. **When developing guidance (RFI #1.a.2) and nomenclature (RFI #3) for AI evaluation, consider the needs and perspectives of both internal *and* external auditors.** While many frameworks and guidance focus exclusively on internal auditors, external auditors are crucial proponents of accountability and often have special concerns surrounding data access and legal protections.

2) Audit execution requires a diverse landscape of audit tooling.

In our recent paper, "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling", we conducted a survey of 390 tools used or intended to help with AI audit work and interviewed 35 employees across 24 tech firms, startups, government agencies, universities, non-profits, and law and consulting firms doing AI audit work.[12] We developed a taxonomy of AI audit tooling (Table 1, Figure 1) to help map the landscape of tools. Our database and taxonomy, which can be viewed at tools.auditing-ai.com, includes multiple tools for identifying impacts, documentation, benchmarking, gathering human feedback, field testing, and other areas mentioned in RFI #1.

[10] Raji, Inioluwa Deborah, et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.
[11] Raji, Inioluwa Deborah, and Joy Buolamwini. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products" *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019.
[12] Ojewale, Victor et al., "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling" (Forthcoming, 2024), https://rbsteed.com/papers/accountability-infrastructure. Referenced figures and tables are reproduced here.

Stage	Categories *(Subcategories)*	*N*	Purpose	Examples
Harms Discovery	Education / Awareness *(community education, visioning)*, Incident Reporting *(incident databases, intake forms, bug bounties, hotlines)*, Target Identification *(algorithm visibility)*	45	Help auditors identify and prioritize audit targets and harms to investigate.	ACLU Wa.'s Algorithm Equity Toolkit, AI Incident Database, Algorithm Tips
Standards Id. & Mgmt.	Goal Articulation *(principle statements, standards formulation)*, Self-Assessment *(checklists, grading)*, Documentation *(single stage, continuous, licenses)*, Regulatory Awareness *(discovery, monitoring)*, Methods Design, Participatory Standards-Setting	194	Help auditors identify and formulate principles and norms to guide their investigations.	AI-RFX Procurement Framework, Microsoft's AI Fairness Checklist, Model Cards [67], Queensland's Community Engagement Toolkit [76], Community Jury
Transparency Infrastructure	Structured/API Access, Secure & Private Sharing *(federated learning)*, Model/Data Exchange	13	Help auditors interact with and analyze proprietary information about the data or model with centralized infrastructure.	NIST's Face Recognition Vendor Test [70], Google AI Test Kitchen [107], Airbnb's Project Lighthouse [4]
Data Collection	Field Data Collection *(scraping, donation, interviews/surveys, compelled disclosure)*, Bot Deployment, Simulation	43	Help auditors collect information about a model's interactions with its subjects.	Mozilla's YouTube Regrets [68], Tracking Exposed, Selenium, Meta's Web-Enabled Simulation [3]
Performance Analysis	Accuracy Evaluation *(A/B testing, benchmarks, adversarial testing, monitoring)*, Explainability *(models, training data)*, Fairness, Qualitative Analysis	126	Help auditors evaluate and explain model behavior through the calculation of performance metrics.	Weights & Biases, Meta's DynaBench, Foolbox, Fairlearn, IBM's AI Fairness 360, Hugging Face's ROOTS [74], Google PAIR's Language Interpretability Tool
Advocacy	Resistance, Community Spaces, Legal Search, Organizing	14	Help organize community action and other accountability measures in response to discovered harms.	Gigbox, Para, Adnauseam, Benefits Tech Advocacy Hub
Audit Communication	Dataset Visualization	2	Help auditors communicate the results of an audit to a broader audience.	Google PAIR's FACETS

Table 1: High-level description of the tool taxonomy categories. Reproduced from Ojewale et al. (2024).[13] Visit https://tools.auditing-ai.com/ for an interactive visualization and links to the full database.

Studying and developing tools for harms discovery, audit communication, and advocacy. In our survey, we found many tools related to evaluation standards, data collection, and performance analysis, including many resources that may be useful to NIST. However, we found six times as many tools for evaluation as we found tools built to help frame and operationalize effective AI evaluations, including tools for harms discovery (for identifying and prioritizing audit targets), audit communication, and advocacy (Figure 1). These areas of audit work—and audit tool development—are equally important for accountability but often under-resourced and neglected.

[13] *Ibid.*

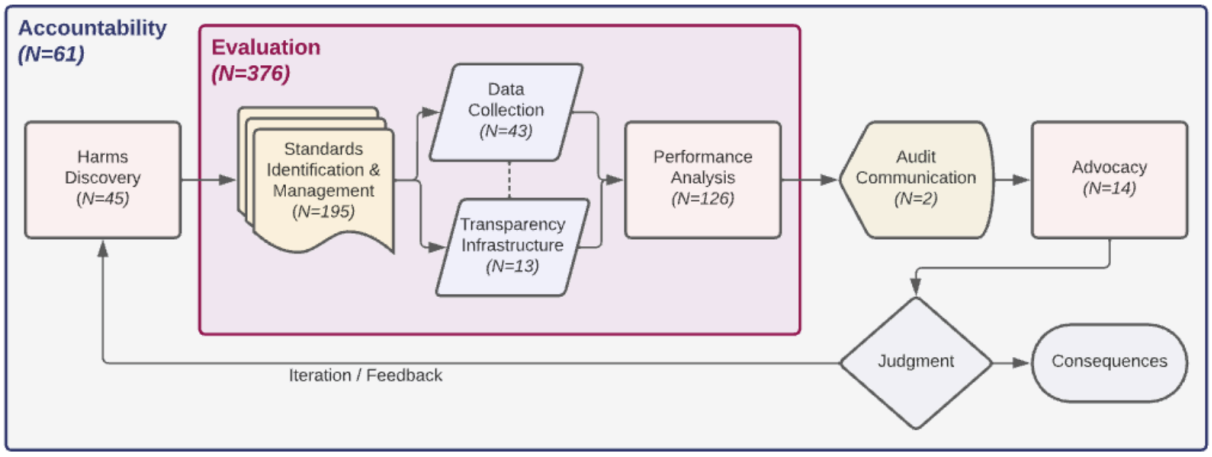

Figure 1: Stages of the tool-supported audit process surfaced in our survey of AI audit tooling. We taxonomize tools by the stage of the AI audit process in which they are meant to be used. Reproduced from Ojewale et al. (2024).[14]

Moving beyond ad hoc toolkits towards common infrastructure. Furthermore, the auditors we interviewed frequently expressed concerns around data access and customization, and multiple audit practitioners expressed the need to move beyond the ad hoc, decentralized array of toolkits currently available and towards resources for common infrastructure. Only a few of the tools we found were supported and maintained by established communities (such as AI Fairness 360, originally developed by IBM and now incubated by the Linux Foundation). Most were supported by only a few independent developers. Participants ran into challenges operationalizing these tools and ensuring their rigor & integrity.

Empowering external auditors. Many tools for evaluation assume easy access to high-quality data, but the external auditors we interviewed struggled to obtain reliable information about AI systems and their behavior. One participant wished for an "inspectability API" (P7), featuring a standardized interface that enables researchers to engage with online platforms to assess diverse treatment, misinformation, and other algorithmic harms by testing various profiles, geographies, and other essential variables. Others suggested centralized databases or archives with necessary information about a system that could be freely used by the public. Participants also wished for efforts to share educational and training resources, open-source audit tools, provide decision-support frameworks to help auditors select appropriate tools for their use cases, and provide resources that help institutionalize AI audit tool maintenance.

[14] *Ibid*.

Recommendations for NIST

1. **Support efforts to centrally record and make available the results of AI audits and descriptions of AI systems and training data (RFI #1.a.2, #3).** In addition to empowering external auditors, making audit work more transparent, and increasing accountability, developing a centralized registry of audit reports can help foster shared norms and standards for effective evaluation and provide a valuable guide for both internal and external audit practitioners.
2. **Broaden the scope of guidance (RFI #1.a.2) and investment to include and support tools and methods for harms discovery, audit communication, advocacy, and other equally important components of effective evaluation.** Our database (tools.auditing-ai.com) contains many examples that may serve as a useful starting point.
3. **Support efforts to develop and *maintain* vetted, open source repositories for AI audit tooling (RFI #1.a.2).** In addition to supporting tool development *and maintenance*, NIST should also contribute to standards and guidelines for the quality of AI tools, particularly methods for explainability (such as SHAP) that are prone to misuse.[15]
4. **Continue to develop and support workshops, forums, and other community platforms for AI audit practitioners (RFI #3)** to communicate and share resources. Platforms such as NIST's AI Metrology series provide a valuable space for practitioners to share experiences, methodologies, and challenges, contributing to a collective pool of knowledge within the broader auditing community. Community-based auditing methods play a vital role in uncovering potential biases and test cases that may be missed in controlled settings—fostering community awareness, mobilization, and public accountability. NIST should consider expanding these programs by convening physical or online interdisciplinary spaces for the audit practitioner community.

3) Expand the scope of evaluation beyond products, models, and algorithms.

In our recent paper, "AI Auditing: The Broken Bus on the Road to AI Accountability", we conducted a survey of 341 AI audit studies published in academic conferences and outlets, supplemented with audit reports, websites, and other documents from AI auditors at news organizations, civil society non-profits, law and consulting firms, regulatory agencies, and large corporations.[16] We found that while most audits of AI systems focused on the behavior of specific machine learning models or AI products, a

[15] I. Elizabeth Kumar et al., "Problems with Shapley-Value-Based Explanations as Feature Importance Measures," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, ICML'20 (JMLR.org, 2020), 5491–5500.
[16] Birhane et al., "AI Auditing: The Broken Bus on the Road to AI Accountability," 2024.

much smaller portion examined the *datasets* used to power AI auditing. And even fewer expanded their scope to consider the *ecosystem* of communities and sociotechnical environments impacted by or involved with AI models and products.[17]

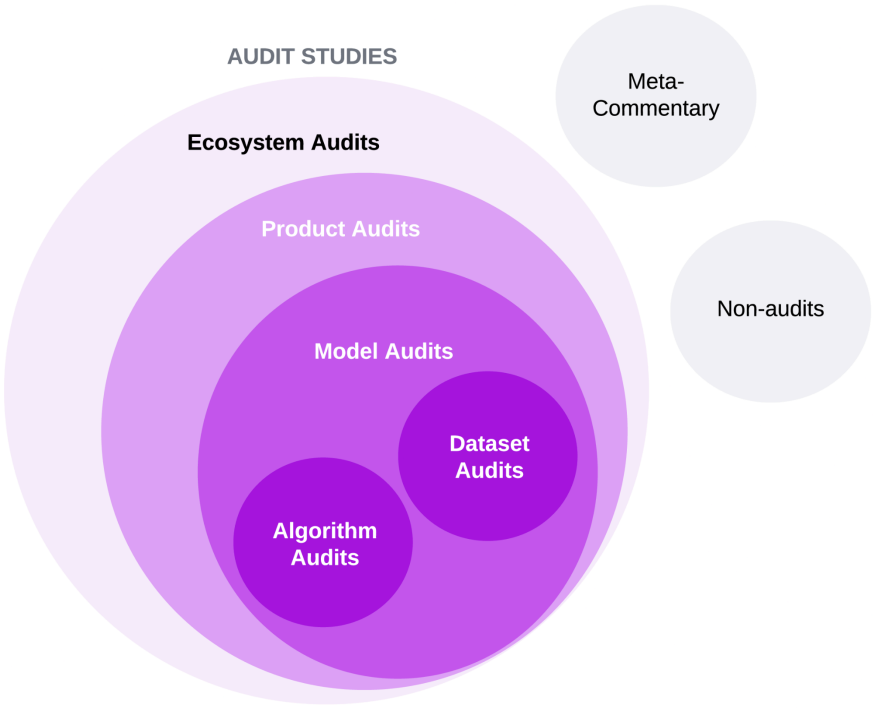

Figure 2: Kinds of audit studies. Reproduced from Birhane et al. (2024).[18]

Auditing ecosystems. Including affected communities and social context is critical to the design and execution of effective evaluations of AI systems, as many AI auditing scholars have argued.[19] We use the term *ecosystem audit* to refer to the idea of examining or evaluating the broader sociotechnical environment that sustains AI (as opposed to directly examining algorithms or AI products), including critical background components to an AI system's operation (such as data labor) and communities that might be impacted as a result of AI deployment. One example of an ecosystem audit is Brown et al., who evaluated child welfare service algorithms through interactive

[17] *Ibid.*, Fig. 2.
[18] *Ibid*.
[19] Evani Radiya-Dixit and Gina Neff, "A Sociotechnical Audit: Assessing Police Use of Facial Recognition," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (New York, NY, USA: Association for Computing Machinery, 2023), 1334–46, https://doi.org/10.1145/3593013.3594084; Michelle S. Lam et al., "Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising," *Proceedings of the ACM on Human-Computer Interaction* 7, no. CSCW2 (October 4, 2023): 360:1-360:37, https://doi.org/10.1145/3610209.

workshops with affected front-line service providers and their families.[20] Ecosystem audits often reveal broader social issues beyond bad AI behavior, including environmental costs,[21] surveillance harms,[22] and labor concerns.[23] In our review, these studies used a broader range of methods (including more qualitative and participatory approaches) and more often produced proposals for specific institutional or policy reforms to address more concretely defined harms.[24]

Auditing datasets. Likewise, some of the most impactful audit studies to date focus on the datasets used to construct AI systems rather than the models or algorithms themselves. Evaluations of internet corpora used to train OpenAI's GPT models were cited in the New York Times lawsuit against Microsoft and OpenAI.[25] And the massive image dataset LAION-5B, used to train popular image generation models, was recently taken down after studies revealed it contained child sexual abuse material (CSAM), pornography, hate content, and stereotypical imagery.[26]

Access, in particular, was repeatedly noted as one of the most common concerns amongst our participants. Few of the large scale image, language, and multimodal datasets used to train commercial generative AI systems are open sourced, and auditors and the general public do not have access to the training sets behind most of the prominent generative AI systems. External auditors often encounter limitations when accessing proprietary systems or data, which leads them to rely on tools like APIs, data scrapers, and crowdsourcing methods. However, these approaches have vulnerabilities, and organizations can restrict access if audit findings are unfavorable. Relying on vendors runs the risk of vendors being able to choose to select particular data for

[20] Anna Brown et al., "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19 (New York, NY, USA: Association for Computing Machinery, 2019), 1–12, https://doi.org/10.1145/3290605.3300271.
[21] Bogdana Rakova and Roel Dobbe, "Algorithms as Social-Ecological-Technological Systems: An Environmental Justice Lens on Algorithmic Audits," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (New York, NY, USA: Association for Computing Machinery, 2023), 491, https://doi.org/10.1145/3593013.3594014.
[22] Radiya-Dixit and Neff, "A Sociotechnical Audit: Assessing Police Use of Facial Recognition," 2023.
[23] Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven: Yale University Press, 2021).
[24] Birhane et al., "AI Auditing: The Broken Bus on the Road to AI Accountability," 2024.
[25] The New York Times Company v. Microsoft and Open AI, No. 1:23-cv-11195, accessed January 25, 2024; Jesse Dodge et al., "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2021, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021), 1286–1305, https://doi.org/10.18653/v1/2021.emnlp-main.98.
[26] David Thiel, "Identifying and Eliminating CSAM in Generative ML Training Data and Models," 2023, https://doi.org/10.25740/kh752sm9123; Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes" (arXiv, October 5, 2021), https://doi.org/10.48550/arXiv.2110.01963; Abeba Birhane et al., "Into the LAIONs Den: Investigating Hate in Multimodal Datasets" (arXiv, November 6, 2023), https://openreview.net/forum?id=6URyQ9QhYv.

favorable audit results, or they may impose requirements on auditors with inside access to limit how the system can be audited and what is allowed to be released to the public. Because of this, our participants frequently struggled to create their own custom-built tools, which takes effort and presents challenges to standardization.

To foster community development and standardization, shared resources such as a common "watering hole" for auditors to access necessary information about the system and tools for evaluation are likely to alleviate independent, ad-hoc tool development and evaluation and therefore enable the community to become more aware of each other's work, which could otherwise be limited by disciplinary or geographical differences.

Utilizing a wider range of audit methodologies. In our survey, we also looked beyond academia. Some of the most impactful AI audit work was conducted in other domains—for example, by journalists, regulators, and civil society non-profits. In particular, while quantitative methods were very common in AI audit papers published at computing venues, auditors outside academia more frequently used qualitative techniques such as investigative reporting, document review, or stakeholder consultation.[27] Guidance should consider that quantitative methods are not the only or most effective way to evaluate AI systems.

Increasing impact with specificity. Similarly, our survey reiterated previous findings[28] that the studies most likely to result in material improvements and mitigations are the audits that have specific targets, objectives, and intended responses. Audits targeting specific systems and using specific criteria for accountability—such as those conducted by the ACLU, the Markup, or the ICO—were generally more effective than more general evaluations not aimed at specific harms or systems.[29] Especially for generative AI systems, which may have especially diverse use cases and impacts, guidance should emphasize the importance of specificity in the targets and objectives of evaluation.

Recommendations for NIST

1. **Guidance on auditing AI systems (RFI #1.a.2, #3) should include and encourage *dataset audits* and *ecosystem audits*,** in addition to the traditional model/product evaluations that dominated our review of contemporary audit studies.

[27] Birhane et al., "AI Auditing: The Broken Bus on the Road to AI Accountability," 2024.
[28] Inioluwa Deborah Raji and Joy Buolamwini, "Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," *Communications of the ACM* 66, no. 1 (December 20, 2022): 101–8, https://doi.org/10.1145/3571151.
[29] *Ibid*.

2. **Guidance should also incorporate a diverse range of methods for auditing AI systems**, including qualitative methods involving investigative reporting, documents review, and stakeholder consultation.
3. **Guidance should encourage context-specificity in auditing methods** and require specificity in the statement of audit objectives and audit targets.

Conclusion

Our studies of the practice of AI auditing suggest that it is essential to recalibrate the focus of audits, especially in the context of generative AI. This recalibration should underscore the overarching goal of *accountability*. Instead of merely evaluating systems, audits should also develop and implement strategies to address issues and ensure that relevant stakeholders are held accountable for the outcomes of these audits.

The entire AI pipeline and auditor-auditee relationship is permeated by complex relational dynamics and uneven power asymmetries. While those that develop and deploy AI systems wield significant power, influence, and resources, those at the receiving end often have little power or agency. Subsequently, acknowledging these factors is crucial for audits to have a meaningful impact. Redressing this imbalance involves fostering equitable and collaborative relationships, thereby ensuring that the auditing process is inclusive and respectful of all voices, particularly those who are most impacted by these algorithms. This also emphasizes the place of participatory methods in ensuring that audits are not just technical assessments but also consider the social, ethical, and real-world implications of algorithms.

The issue of funding for AI audit work is also critical. Adequate funding — for tools, community spaces, frameworks, standards, and for auditors themselves — ensures that audits are thorough and comprehensive, covering all necessary aspects of these complex systems. It is important to advocate for resources that match the complexity and societal importance of these algorithms, ensuring that financial constraints do not compromise the quality of audits.

Finally, recognizing the limits of audits is crucial for a balanced understanding of their role in algorithm governance. While audits are a necessary tool for ensuring accountability and transparency, they are not a complete solution — they are just a step in the direction of creating socially responsible systems.