# **Mozilla OAT NTIA Comment**

The Mozilla Open Source Audit Tooling (OAT) project aims to identify the resources and tools that can support auditors of all types to analyze AI systems and push towards a thorough and consequential scrutiny of these systems. As part of this project, we analyzed over 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 100 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, with the support of research assistant Victor Ojewale.

This project complements other related initiatives supported by the Mozilla Foundation, including the Mozilla Technology Fund's support of the development of algorithm audit tools and the Data Futures Lab's support of crowdsourced data donation models.

This research in progress also complements prior work on institutional design requirements for an effective external audit ecosystem[1], practical strategies for internal[2] and external[3] auditing investigations, and commentary on the nature of those identifying as algorithm auditors[4].

Several of the findings in this work are relevant to the NTIA call for comment.

## 1) There are two populations of "algorithm auditors", each with their own distinct motivations, skill sets and challenges.

The definition of an 'AI audit' is made ambiguous by the fact that those identifying as "algorithm auditors" effectively encompass two unique populations. In our report, we use the taxonomy of "internal" and "external" auditors to distinguish the range of participants in the investigations on algorithmic harms. We consider internal auditors, typically with a contractual relationship with the audit target, as those who conduct an independent review of the development and deployment of the products in use. External auditors are considered fully independent entities, unlinked to the audit target, that engage in investigations, typically on behalf of the interests of represented constituents. The contexts and goals of the two camps differ meaningfully. Internal auditors typically operate under professional obligation and processes are designed for those that seek to validate procedural expectations, aim to minimize liability and test for compliance to

[1] Raji, Inioluwa Deborah, et al. "Outsider oversight: Designing a third party audit ecosystem for ai governance." *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022.

[2] 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.

[3] Bandy, Jack. "Problematic machine behavior: A systematic literature review of algorithm audits." *Proceedings of the acm on human-computer interaction* 5.CSCW1 (2021): 1-34.

[4] Costanza-Chock, Sasha, Inioluwa Deborah Raji, and Joy Buolamwini. "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem." *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.

AI principles and legal constraints[5]. 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[6]. Each of these groups exhibits unique motivations, skill sets, and obstacles that arise from varying responsibilities within auditing procedures and the objectives they aim to achieve.

It is not enough to simply conduct an audit; the outcomes of the audit must be recognized and acted upon, and organizations held responsible for addressing any identified harms. The NTIA defines an audit as "an external review at a point in time against accepted benchmarks… may be conducted by internal or external reviewers". Although it is unclear what 'external' means in this context, this definition aligns with our understanding of an audit as an *independent* review conducted by internal or external actors, where independence signifies meaningful separation between the auditor and those developing the system being audited. However, we further distinguish an ***"audit" study*** from other algorithmic assessments by anchoring the former to a focus on **concrete evaluations with the expectation of accountability** rather than a higher level process of reflection and system analysis. This definition follows a tradition of similar research investigations in the social science context[7], where such audit studies directly informed advocacy for improved social justice outcomes for the impacted parties. A thorough audit should thus not only build trust for external stakeholders as the NTA states, but should also include driving improvements within an organization's internal operations, or inciting other consequences in response to the audit result, in order to hold audit targets accountable for addressing the outcomes of the audit.

Internal auditors are responsible for ensuring compliance. In sectors where audits are mandatory, a specialized industry of professional auditors emerges to fulfill this demand. These auditors are typically appointed within organizations and are tasked with conducting thorough assessments that ensure compliance with relevant standards and regulations, but can also involve hired consultants. Their role is not limited to oversight — they actively engage in signing contracts with audit targets that grant them full access to the necessary systems and information required to perform, and set the terms of the audit. The main objectives of internal auditors revolve around compliance, which entails aligning internal procedures with anything from internal principles or objectives; external mandates; or industry standards and best practices. Motivations behind this need for compliance can vary. On one hand, the aim may be to help various stakeholders such as end users and regulators build confidence and "trustworthiness" in the system by demonstrating adherence to external standards and regulations. This is particularly important in industries where public trust is dominant and external stakeholders

[5] 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.

[6] 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.

[7] Vecchione, Briana, Karen Levy, and Solon Barocas. "Algorithmic auditing and social justice: Lessons from the history of audit studies." *Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021. 1-9.

need assurance that the system works reliably and ethically. On the other hand, compliance may be primarily concerned with aligning internal processes with established best practices in the field, which helps organizations mitigate risks.

External auditors encompass a distinct group of researchers and practitioners who engage in voluntary opt-in investigations that explore their own inquiries around algorithmic deployments. Unlike internal auditors, who primarily focus on compliance and internal processes, external auditors approach their work with goals more closely aligned with advocacy and social justice. The approaches taken by external auditors play a crucial role in examining an algorithm's potential harms as well as its social implications. Historically, external auditors have played a crucial role in unearthing evidence of AI harms and exposing avenues of risk.The organizations that have played this role of "external investigators" have been vastly diverse, ranging from law firms to journalists to civil society to academic researchers. Some of the most impactful investigations into the harms perpetuated by these platforms today have been conducted by these groups, which effectively operate as public interest researchers with an advocacy focus and interest. External auditors tend to understand that communities possess unique perspectives and can generate questions that go beyond what a company or single auditor can do in isolation. These external auditors tend to be aware of the need to embrace more collaborative and participatory approaches that involve the community. Involving participants in the audit development and evaluation process can spur a sense of ownership and empowerment that ultimately promotes a shared commitment to addressing the potential impacts of the algorithm for both individuals and the whole of society. Ultimately, these collective participatory efforts allow external auditors to shift power to the impacted communities, thereby extending the scope of the audit and leading to more comprehensive and robust assessments.

### 2) These two populations of "algorithm auditors" have differing needs in terms of tools and infrastructure for support in execution.

The execution of an audit or investigation study can be incredibly difficult. Through our survey of audit tools and case studies, alongside interviews with practitioners, we have already begun to identify meaningful pain points in the audit process and areas of ongoing tool development to address these challenges for both internal and external audits. The execution of "external audit" studies in particular is incredibly diverse and faces unique difficulties. As many researchers are not necessarily capable of participating in tool development themselves, it will be crucial for there to exist a robust ecosystem of audit study tools that specifically lower the barrier to participation in this public-interest research. These tools must be supported in addition to efforts facilitating internal audit processes, where standardization, measurement challenges and the complications of multi-stakeholder communication lead to their own set of execution challenges.

Although this is a work in progress, we lay out the major activities and pain points we have identified below. Many of the tools and methodologies we examined are useful to both external and internal auditors, but we notice several preliminary differences in the kinds of tools created by and designed for each group of auditors. In addition to these differences, algorithmic products are quite varied, and the subject of the audit could range from recommendation systems on online platforms, automated decision systems (ADS) or foundation models. Each type of audit subject naturally involves differing types of methodologies that also introduce novel challenges. For instance, an external audit researcher for an online platform audit study will not have difficulty identifying the audit target (typically a large online platform) but may struggle with data access and generating reproducible evaluations, robust to independent changes on the platform. On the other hand, an external auditor analyzing an ADS system may have much difficulty identifying the vendor of the system in the first place, as many of these systems are typically not visible to impacted population members.

**External auditors**, who by definition may not have access to privileged information and existing data about system behavior, particularly value tools for **harms discovery**, **data collection**, and **data and model access mediation** (structured transparency). In our survey, existing tools for this purpose were developed most often by non-profits and other third parties and intended most often for external practitioners. These tools help keep the door open for external audit researchers, both by identifying targets for audits, exploring potential harms, and providing vital evidence for investigations.

- **Harms Discovery:** One challenge in algorithm auditing is identifying which targets to audit and how to understand what to audit for in order to meaningfully protect vulnerable populations, especially for external auditors who are concerned about harms but may not be aware of proprietary or otherwise hidden AI systems or their possible impacts. Tools for harms discovery help identify and select audit targets in addition to supporting the identification, characterization, and prioritization of algorithmic harm experiences to investigate. This category includes tools for Education & Awareness (to engage and involve community stakeholders in articulating harms), Incident Reporting (to intake and solicit public reports of algorithmic harms, including bug bounties), and Target Identification (to uncover deployed systems and make them visible to external practitioners). This category is relatively neglected, compared to other kinds of tooling — there are only a few centralized resources or processes for identifying and pursuing reports of algorithmic harm, typically maintained by one or two individuals or non-profits.

- **Data Collection:** The absence of effective tools for data collection presents a significant challenge for auditors aiming to gather empirical evidence as part of an algorithmic audit, especially for auditors who do not already have access to data collected by model operators. These tools are essential in gathering information about the interactions between a model and its subjects, which is a vital component of the audit process, and often include new and relevant information not routinely collected by model operators. This category includes tools for Field Data Collection — from Data Donation, Data Scraping, and qualitative Interviewing to tools for Compelled Transparency (e.g. tools that facilitate FOIA requests) — as well as tools for Simulation and Bot Deployment (for sock puppet auditing) to test systems in controlled artificial or semi-artificial interactions. Unfortunately, these tools may violate platform terms of service and can be impeded if model operators take technical or legal action.

- **Data and Model Access Mediation:** When the corporations operating large AI systems are unwilling or unable to release relevant documentation and other evidence publicly, these tools provide organized, comprehensive, and centralized infrastructure and are particularly important for external auditors to hold the machine learning community responsible for careful data distribution. Instead of simply uploading data to a cloud-supported drive folder, the rise of these tools helps promote responsible and accessible use of valuable information through concrete and secure infrastructure. This category includes Application Programming Interfaces (APIs) (for interacting with models and live systems at scale), Secure Databases (for sharing sensitive, audit-relevant data securely and privately), tools for Data Pooling (for aggregating data across organizations or silos), and tools for Model/Data Exchange (for collating donated data into a central trust). While some instances of these tools currently exist, they require voluntary investment by model operators, and many external auditors feel that currently available tooling is inadequate for fully investigating deployed systems.

Tools created by **internal auditors**, on the other hand, focus proportionately more on quantitative **performance analysis** for models in operation**.** Tools for **standards identification & management** were also mostly intended for internal audits**,** but most of the standards frameworks and principle statements we reviewed were created by third-party non-profits or government agencies. While some organizations conducting internal audits published information about their standards publicly, others maintained internal but proprietary documents. Still others may not have clear standards for their internal audits at all.

- **Standards Identification and Management:** Auditors — internal and external — require additional tools to identify and formulate principles and norms to guide their investigations. This category includes tools for Goal Articulation (for broad principles), Checklists (for specific process requirements), Documentation (including all stages of model development and deployment), and Regulatory Awareness (for discovering and monitoring relevant legal requirements for AI products). The limited development of these tools impedes auditors from establishing robust frameworks to conduct their assessments and may result in inconsistencies and ambiguity in their audit processes and objectives. Additionally, the opacity of internal audit frameworks makes it difficult to assess and communicate the efficacy of internal audits.

- **Performance Analysis:** The availability of tools for performance analysis presents a notable pain point for auditors as they attempt to evaluate and explain model behavior by calculating performance metrics. This category includes tools for Fairness Evaluation (for determining whether the target system treats groups or individuals unequally), Accuracy Evaluation (for revealing general capabilities and limitations), Benchmarking (for establishing standardized performance characteristics), A/B Testing (for comparing models in production), Adversarial Testing, Model Monitoring, Model Explainability, and Training Dataset Exploration. This popular category of tools also emerges as one of the most prevalent areas of practitioner concern in our survey — while many tools exist in this space, there is a pressing need for robust, vetted tools and methodologies that facilitate accurate evaluation and explanation of algorithmic performance.

We also identified several nascent categories of tools that could aid both internal and external audit practitioners, including tools for **audit communication** and **advocacy**. These are underdeveloped, important areas of audit tooling often ignored in discussions of audit resources and new development efforts.

- **Audit Communication:** A limited ability to effectively communicate audit results to a broader audience hinders outside parties from actively participating in the identification, discussion, and advocacy for responsible and transparent AI auditing practices, and makes it difficult for internal auditors to garner trust and credibility. This emerging category of tools may include tools for Community Engagement (from surveys to storytelling), Dataset & Other Visualizations (for communicating results in a more accessible way), Media & Press Communication (for disseminating results), and Audit Transparency & Reporting (for centralizing and standardizing audit reports).

- **Advocacy**: Lack of adequate tools for reporting and community action poses a challenge for the broader community to stay informed about existing audits, report emerging audit reports, and effectively organize accountability measures. Even internal auditors value tools that allow them to advocate for changes within their organizations. This emerging category includes Legal Case Databases & AI Audit Case Databases (for identifying relevant legal and methodological precedent), Audit Report Platforms (to notify journalists and other practitioners of new reports/publications), and Collective Organization & Action Tools (such as community spaces for practitioners to convene, organize, and collaborate).

3) **These two populations of "algorithm auditors" require different policy interventions. Any meaningful long term regulatory ecosystem should account for both groups.**

As internal auditors theoretically have full access to the entire audit target organization and set of engineering artifacts, the most common failure mode that they may experience is to publish an assessment that falls short of capturing meaningful information regarding external expectations for the deployed system. External auditors face a similar failure mode in audit quality but for different reasons, mostly related to challenges with data access -- in other words, not having enough information to do a proper and accurate assessment of the system. As a result, the function of policy for both populations of "algorithm auditors" differs in certain ways, and overlaps in others.

It is noteworthy that it is likely that both groups also respond to differing incentive structures. Internal auditors are largely the byproduct of some perceived or anticipated compliance need - if a company feels the need to pre-empt or adhere to given standards or external expectations for product vetting, then there is some acknowledged need for internal audit teams to be established. However, without these policy measures, there is little direct incentive for corporations to hire and maintain internal audit teams or consultants, except in cases of extraordinary public pressure or organizational foresight. In a similar way, advocacy work requiring external investigations is not feasible without obtaining adequate protection against corporate retaliation to minimize the risk to  advocates hoping to get engaged in this work.

Much of the regulatory requirements for internal auditors or professional audit actors is an **enforcement of some degree of visibility or oversight** on their internal assessment processes and outcomes, which currently remain relatively obscure to external stakeholders, including regulators and the public. This means some degree of internal control or process oversight to ensure adherence to best practices in audit process and methodology, as well as judging conflict of interest to maintain a reasonable degree of independence in evaluations leveraged in accountability processes. This could also involve making visible internal audit outcomes, either through the direct publication or registry of audit results or by requiring the publication of internal audit reports. Due to the lack of commercial incentives, regulation will need to play a role in the external communication of internal audit processes for oversight and scrutiny, including requiring the open source distribution of documentation templates and other forms of internal audit infrastructure, including possibly open sourcing details of data access APIs.

Policy should also hold the door open for external auditors, who effectively operate as voluntary researchers and investigators of deployed algorithmic systems. Regulators can support these external actors in various ways: particularly, in terms of their information needs as it relates to **data access** and **target identification**. For access, external auditors need safe harbors against retaliation for the publication of unfavorable results and custom tooling for data collection. As mentioned previously, many existing tools for data collection (ie. corporate mediated APIs) are insufficient for the needs of these external investigators. Thinking through structured mediation through regulators and support for external research access will be a critical intervention to support long term external accountability measures. Regarding target identification, regulators can mandate and maintain AI registries for deployments within government of various algorithmic products; and require active notice of AI use to impacted populations in administrative processes such as hiring. Additionally, information should be provided that is helpful for external parties to produce reliable investigations. Especially for platform audits, there needs to be required external communication from audit targets regarding meaningful changes to the product (ie. algorithm updates, data sample details) that may impact the reproducibility or findings of an audit evaluation.

External audit studies or investigations can take on many forms and include a range of methodologies. These studies can be qualitative or quantitative. As a result, an informative “data” inquiry and release does not necessarily involve records of mandatory platform data disclosures on user profile data or algorithmic scores. It could also include the release of other forms of helpful information that could illuminate other aspects of platform governance, including the release of internal documentation, the release or opportunity to gain access to interviewing key internal stakeholders, and other forms of qualitative evidence. Audit studies can also go beyond the scope of algorithm analysis. For instance, many audit studies also involve investigating user actions, data practices, and broader-level societal or institutional effects of platform engineering decision-making.

In addition to the above points, in both cases, some policy directions that will be critical for future development also involve the following key policy areas:

- **Credentialism:** The notion of accreditation in the field is still a debated topic; at the moment, there is no formal centralized credentialing of those operating in this capacity[8]. It will be important that both industry players and voluntary accountability actors adhere to some basic expectations around auditor conduct and audit practitioner best practices. Thus, in both cases, an audit oversight board should be considered, in order to inspect credibility, independence and methodological rigor. Ideally, auditors are credentialed by regulators and not companies - even mandatory internal auditors are ideally paid and selected by government actors and not industry to maintain independence. Similarly, “external investigators” can take on all kinds of forms of organizations, ensuring minimal compliance to certain expectations with regard to research ethics or formalizations ensures the responsible handling of data and overall integrity during the evaluation process.

- **Impact:** Both professional audit practitioners and advocacy-oriented external investigators seek to have some influence in holding audit targets accountable for their consequential decision-making regarding the algorithmic product. Ideally, there is some mechanism for reporting audit outcomes directly to a regulator (ie. through an audit report registry) or for mandatory responses from corporations to the audit report. In addition, we find that many external investigators value the opportunity to influence or impact downstream regulatory action. In addition to access, there should be guaranteed support for ensuring that research outcomes can be directly communicated to and acted upon by key decision-makers at audited corporations and regulatory organizations.

- **Resources and development:** Overall, audit methodology is very nascent and the processes are high cost -- investment will be required to evolve the ecosystem into one that can sustain accountability needs, both internally and externally. Government support for research in the area of developing adequate evaluation and impact assessment techniques is required, in addition to the support and development of the necessary audit infrastructure and tooling to make the execution of audits feasible.

[8] Costanza-Chock, Sasha, Inioluwa Deborah Raji, and Joy Buolamwini. "Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem." *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.