Should AI Auditors Be Audited?

Challenges and Paths for Meta-Auditing Artificial Intelligence

Marcelo Pasetti[1]

Abstract: *The growing use of artificial intelligence (AI) in public administration, corporate governance, and high-risk sectors such as health, justice, public safety, and credit, which are recognized as domains that affect fundamental rights, human dignity, and access to essential services, has intensified debates about the need for reliable, transparent, and independent algorithmic audits. Although these audits are often presented as instruments of transparency and accountability, the article demonstrates that the current audit ecosystem remains marked by technical deficiencies, conflicts of interest, and insufficient oversight of the auditors themselves. Based on a qualitative and normative approach, grounded in a documentary analysis of recent international regulations, institutional technical guidelines, and specialized scholarly literature, the study examines the role of audits as governance mechanisms and the risks arising from regulatory capture, private self-regulation, and institutional opacity. Finally, it advocates for the creation of multilevel audit ecosystems, anchored in public or hybrid meta-audit structures, social participation, and shared standards of integrity, transparency, and democratic legitimacy.*

Keywords: *Algorithmic auditing; Artificial intelligence; Meta-auditing; Epistemic justice; Accountability*

1 Acknowledgements – This research is part of the RAIES (Rede de Inteligência Artificial Ética e Segura), supported by FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul).

Introduction

The rapid expansion of the use of Artificial Intelligence (AI) systems in public and private sectors has generated growing concerns about the risks of algorithmic bias, decision-making opacity, and structural inequality in the distribution of their impacts. AI tools are already used in sensitive domains such as recommendation systems, health, credit, and criminal justice, often without adequate transparency regarding the operational criteria and responsibilities involved (Bandy 2021; Mittelstadt 2016). In this context, algorithmic audits have been consolidated as technical and regulatory mechanisms for verifying automated systems' compliance, fairness, and legality. However, as Jack Stilgoe (2023) warns, public trust in AI is not based solely on technical tests; it requires a reflexive approach to the role of auditors who hold epistemic power in evaluation processes. This raises broader political concerns in liberal-democratic systems, where insufficient oversight may generate adverse effects on citizens' rights and public accountability.

As auditing practices become institutionalized, a central question arises: **should AI auditors be audited?** The growing professionalization of the sector is illustrated by initiatives such as the Ada Lovelace Institute's *Code & Conduct: How to Create Third-Party Auditing Regimes for AI* and the development of technical standards by the IEEE Standards Association. Even so, the scenario remains in formation, marked by asymmetries of epistemic and political power and the absence of structured mechanisms for supervising the auditors of algorithmic systems themselves. As demonstrated in the empirical mapping conducted by Costanza-Chock et al. (2022), many audits are carried out by internal teams or contracted directly by the companies responsible for the audited systems, compromising the results' impartiality. The lack of consolidated standards and robust regulation can turn audits into mere reputational validation instruments rather than effective accountability and transparency mechanisms.

This research investigates the ecosystem of algorithmic audits with a specific focus on the lack of oversight over AI auditors, to understand the main epistemic, regulatory, and political risks arising from the lack of institutional accountability. Although the field of auditing is developing, this study adopts the central hypothesis that the effectiveness of audits requires the construction of permanent meta-audit structures capable of ensuring the integrity of the assessment process through independent oversight, systematic review, and public transparency. Without such mechanisms, audits risk reinforcing informa-

tional inequalities and legitimizing technocratic practices without promoting true accountability.

This debate takes on relevance in the European context, with the entry into force of Regulation (EU) 2024/1689 (AI Act), which establishes a comprehensive regulatory architecture for AI's ethical and safe use. The regulation classifies AI systems by risk level and imposes proportionate obligations, providing for pre-market compliance assessments and post-market monitoring mechanisms. Although it represents an important milestone, the AI Act still lacks clear guidelines on the supervision of auditors and on ensuring the institutional integrity of notified bodies (Mökander et al. 2022). Its emphasis on internal self-assessment procedures by providers, further reduces the effectiveness of enforcement and limits the regulatory model's ability to prevent systemic risks.

In this scenario, the proposal for dynamic audits is gaining strength as a viable alternative to traditional static verification. Maughan et al. (2022) introduce the concept of prediction sensitivity, aimed at continuously auditing counterfactual fairness in deployed classifiers and detecting degradations in fairness over time. This approach reinforces the importance of iterative and adaptive audits, especially in systems subject to operational changes. The effectiveness of this proposal, however, depends on robust institutional structures and public policies that support regular and transparent inspection practices, as discussed in the Ada Lovelace Institute, AI Now Institute, and Open Government Partnership report *Algorithmic Accountability for the Public Sector: Learning from the First Wave of Policy Implementation* (2021).

The methodology of this study is qualitative and analytical, based on a bibliographic and documentary review of recent regulatory, institutional, and scientific sources, highlighting regulatory and institutional developments in the European Union and the United States. The theoretical framework adopts the work of Vecchione, Levy, and Barocas (2021), who interpret auditing as a historical tool of social justice, aimed at accountability and the inclusion of affected groups. This perspective is combined with the approach of Schiff, Kelley, and Camacho Ibáñez (2024), whose analysis of regulatory capture, delegation, and informational asymmetry highlights the institutional and political limits of ethical AI auditing. The integration of these two strands, auditing as a tool of social justice and auditing as an institution vulnerable to capture, provides the basis for assessing both the promises and shortcomings of current audit practices. In doing so, the article situates the discussion within broader debates on accountability, epistemic justice, and democratic governance.

In this article, epistemic justice is understood operationally as the equitable distribution of informational and interpretive power in algorithmic evaluation processes. It is about ensuring that affected groups have real opportunities to participate, challenge, and influence the criteria and results of audits, avoiding asymmetries of voice and authority that produce systematic disadvantages (Costanza-Chock et al., 2022; Vecchione, Levy & Barocas, 2021; Stilgoe, 2023).

The article is divided into two sections. The first addresses the structural and epistemological challenges of the current AI audit ecosystem and exposes technical, institutional, and regulatory flaws. The second section presents normative and institutional pathways for strengthening audit governance, focusing on implementing meta-audits and public accountability mechanisms. The conclusion summarizes the findings and indicates guidelines for building a more ethical, trustworthy, and socially inclusive audit ecosystem.

The Algorithmic Audit Ecosystem and Its Structural Challenges

The growing institutionalization of algorithmic auditing as a governance tool has revealed not only its technical potential but also the institutional and regulatory challenges that undermine its effectiveness. When proposing that AI systems be auditable, the need for stable criteria, independence of evaluators, and effective accountability mechanisms is assumed. As Costanza-Chock et al. (2022) point out, although algorithmic audits have gained prominence as accountability instruments, the field still faces the absence of consolidated methodological standards, transparency in procedures, and effective participation by civil society organizations. They argue that a truly robust audit ecosystem requires technical independence, external oversight, and the active involvement of communities potentially impacted by AI systems. Bandy (2021) notes that, despite the growing importance of algorithmic audits, there is still little clarity about the methods used and the criteria that determine an effective audit, highlighting the need for greater systematization and standardization in the field.

Costanza-Chock et al. (2022) state that the algorithmic audit ecosystem remains marked by institutional fragmentation, self-regulation, and a lack of defined criteria regarding the qualifications of auditors and the quality standards of the audits carried out. This framework forms an asymmetrical environment in which accountability is limited and unequal. The disparities in epistemic

power between private auditors and the individuals directly affected by automated decisions aggravate the lack of systemic accountability. As Vecchione, Levy, and Barocas (2021) argue, audits can cease to fulfill their critical control function and become instruments for legitimizing technically validated but socially unjust practices.

Faced with these challenges, this section examines the technical and political functions of audits (Section 1.1) and the institutional and regulatory obstacles that undermine their legitimacy (Section 1.2). Together, these analyses clarify why the question "should AI auditors be audited?" has become central to contemporary algorithmic governance.

Auditing as an Algorithmic Governance Tool

The rapid advance of AI technologies has created an increasing demand for institutional mechanisms that ensure their deployment is ethical, fair, and lawful. In this context, algorithmic auditing has been consolidated as a strategic tool for the governance of these systems. Although there is no universally accepted definition, recent studies advocate for broadening the methodological scope through the introduction of sociotechnical audits. According to Lam et al. (2023), these audits evaluate algorithmic systems considering both the technical components and the impacts on users, broadening the traditional lens to consider the continuous interaction between algorithms and people.

As described in the report *Code & Conduct: How to create third-party auditing regimes for AI,* by the Ada Lovelace Institute (2024), auditing is understood as a systematic process to evaluate algorithmic systems based on explicit criteria of compliance, performance, and responsibility. The sociotechnical approach proposes expanding algorithmic auditing beyond the technical analysis of computational inputs and outputs, incorporating the social, institutional, and political contexts in which systems operate. According to Lam et al. (2023), this audit model recognizes that algorithms do not exist in isolation but interact with broad sociotechnical ecosystems, in which different forms of power shape systems' functioning and their distributive effects.

Among the main functions attributed to AI audits are: ensuring compliance with legal norms, promoting public trust in automated systems, and preventing algorithmic discrimination. However, there are clear limits to their effectiveness. As Schiff et al. (2024) point out, the challenges arise not from the auditors themselves but from the structural conditions under which audits are conducted. Because many audits are commissioned and financed directly by AI

developers, the process becomes vulnerable to conflicts of interest, which undermines professional independence and weakens the credibility of the results in fragile regulatory contexts. Similarly, Costanza-Chock et al. (2022) point out that many algorithmic audits occur without sufficient transparency, making external verification and reproducibility of results difficult. The authors note the absence of a consolidated consensus on the criteria that define a rigorous audit, a gap that contributes to methodological fragmentation and weakens the reliability of auditing as an effective instrument of accountability.

These limits are intensified in the public sector, where outsourcing automated systems to private providers has become a recurring practice, especially in sensitive areas such as policing, social benefits screening, and administrative decisions. The Electronic Privacy Information Center (EPIC) report documents how delegating public functions to contractors can weaken democratic transparency and hinder institutional accountability, especially when systems fail or produce unfair decisions (Fergusson, 2023).

A notable example is the COMPAS recidivism algorithm in the United States, whose risk assessment model revealed systematic racial disparities in criminal sentencing outcomes (Angwin et al., 2016; Angwin et al., 2016b). Similarly, the Dutch Child Benefits Scandal, examined by researchers from the University of Groningen (Bouwmeester, 2023), exposed how an automated fraud detection system unjustly flagged thousands of families, disproportionately those with dual nationality, resulting in severe financial and social harm. Both cases illustrate how opaque or poorly supervised algorithmic systems, when embedded in public administration, can perpetuate structural discrimination and erode public trust in governmental institutions.

In this sense, algorithmic auditing transcends the technical sphere, becoming a political mechanism that directly influences the legitimacy of automated decisions. The absence of regulations establishing minimum independence standards, methodological transparency, and external supervision can turn the auditing process into an opaque formality, potentially legitimizing technically verified injustices. Stilgoe (2023) emphasizes that evaluating AI technologies requires not only assessing their technical effectiveness but also the public value they generate, underscoring the need for institutional structures that safeguard democratic accountability.

Diagnosis of Systemic Failures and Risks

The consolidation of algorithmic audits as governance instruments encounters persistent obstacles that undermine their effectiveness, impartiality, and democratic legitimacy. These obstacles, widely documented in recent literature, can be categorized into technical, institutional, and regulatory failures. Schiff et al. (2024) show that information asymmetries, conflicts of interest, and a lack of methodological standardization among auditors mark the audit ecosystem. Bandy (2021) points out that most studies on algorithmic audits focus exclusively on quantifiable metrics, neglecting social and structural impacts. Mökander et al. (2022) and Wörsdörfer (2023) point out that the European Union's regulatory model lacks clear and effective mechanisms for supervising auditors, which weakens *institutional accountability.* Diagnosing these structural flaws is fundamental to supporting the debate on the need for independent and continuous meta-audit structures capable of conferring public legitimacy on the regulatory process and preventing institutional capture phenomena.

On a technical level, most audits focus on metrics such as accuracy, recall, and statistical precision of AI systems, but neglect critical aspects related to algorithmic fairness, social impact, and indirect discrimination. As Bandy (2021) further observes, many studies on algorithmic auditing prioritize quantitative approaches, focusing on metrics such as accuracy and unequal impact, while neglecting broader issues of social justice and institutional context, which can limit the identification of systematic biases.

Maughan et al. (2022) propose the concept of *prediction sensitivity* as a metric aimed at continuously auditing counterfactual fairness in classifiers that have already been implemented. This approach seeks to measure whether and how the output of a system would change if the individual being evaluated belonged to a different social group, with all other characteristics remaining constant. By allowing the detection of subtle asymmetries in treatment, even in the explicit absence of sensitive attributes, this metric highlights the inadequacy of one-off audits carried out only during system development. Based on this limitation, the authors argue that algorithmic fairness should be monitored over time, in real contexts of use, which requires an iterative and continuous audit model. In this context, the proposal for so-called dynamic audits gains relevance, as it prioritizes constant monitoring of the performance of systems in operation as the most appropriate strategy for dealing with the distributive and epistemic risks associated with adaptive classifiers.

At the institutional level, AI audits are often conducted by companies hired or financed by the developers or operators of the audited systems, generating a structural conflict of interest. This economic dependence compromises the auditors' independence and undermines their conclusions' credibility. Schiff et al. (2024) warn that, without robust criteria and independent supervision, ethical AI audits can be appropriated by organizational dynamics, serving more as symbolic instruments than effective accountability mechanisms. The lack of methodological standardization among auditing firms aggravates this dynamic. As the Holistic AI report (2023) points out, the absence of internationally consolidated standards for audits of large language models (LLMs) results in divergent approaches between organizations in terms of methodology and evaluation criteria. This disparity can jeopardize audit comparability and make it difficult to form a reliable and interoperable algorithmic enforcement ecosystem.

On a regulatory level, the AI Act represents a significant milestone by establishing requirements proportionate to the risk of AI systems, including obligations for conformity assessment, post-market monitoring, and the designation of notified bodies. These bodies must demonstrate technical competence, independence, and impartiality. However, the regulation gives member states a broad discretion regarding these bodies' designation, notification, and supervision, without providing detailed and uniform mechanisms for external auditing or ongoing public accountability. This gap can compromise the robustness of oversight, especially in sensitive areas. Mökander et al. (2022) have already warned that the proposed European regulation lacks operational guidelines on the supervision of auditors, especially about their institutional independence, the integrity of compliance processes, and the responsibility of notified bodies.

Even after the publication of Delegated Regulation (EU) 2024/1773 by the European Commission, which details criteria on the organizational independence of notified bodies, management of conflicts of interest, and documentation requirements, gaps remain about external oversight and the effective accountability of these entities. The regulatory text does not establish concrete meta-audit mechanisms or impose clear obligations of public transparency on the audit procedures carried out, allowing these bodies to act with a wide margin of autonomy and low exposure to institutional scrutiny. This weakness compromises the effectiveness and legitimacy of the European Union's conformity assessment model, especially in sensitive AI applications.

These technical, institutional, and regulatory failures can be interpreted through the lens of regulatory capture theory, as formulated by Stigler (1971), which posits that regulatory bodies tend to be gradually appropriated by the sectors they are intended to oversee. In domains characterized by high technical complexity and low transparency, such as AI auditing, this capture occurs in diffuse and incremental ways, aligning auditors over time with the values and narratives of the organizations they evaluate. This process undermines critical independence and diminishes the transformative potential of audits, as also noted by Schiff et al. (2024), who warn of the institutional absorption of corporate priorities by the evaluators themselves.

In addition to these distortions, there are the classic problems of the principal–agent relationship, in which the public (as principal) delegates supervisory functions to technically specialized agents (auditors), whose objectives do not always converge with the collective interest, and whose actions usually occur with low visibility and little accountability (Eisenhardt, 1989). As the report *Algorithmic Accountability for the Public Sector* (2021) summarizes, implementing effective algorithmic accountability policies faces structural obstacles, such as institutional fragmentation, the absence of common terminology between public bodies, and the lack of clear legal incentives to support robust and continuous supervision.

This set of factors reveals a fragile algorithmic audit ecosystem, marked by low standardization, conflicts of interest, and the absence of independent oversight mechanisms. The prevalence of symbolic audits and the lack of accountability on the part of auditors undermine the legitimacy of the entire process. These flaws highlight the need to audit the auditors themselves, thereby clarifying the central question of this study: should AI auditors be audited?? Additionally, these challenges underscore the urgent need for a broader normative and institutional reconfiguration of AI auditing practices, grounded in independence, transparency, and democratic accountability.

Institutional, Regulatory and Technical Pathways Towards Trustworthy Audits

Considering the technical, institutional, and regulatory flaws identified in the current algorithmic audit ecosystem, advancing beyond isolated adjustments has become imperative. Consolidating trustworthy audits requires a comprehensive reconfiguration grounded in multi-level governance, articulating legal

guidelines, internationally recognized technical standards, effective accountability mechanisms, and channels for informed public participation. As noted in the *Code & Conduct* report by the Ada Lovelace Institute (2024), the creation of auditable ecosystems requires not only consistent assessment methods, but also legitimized institutional structures that are open to public scrutiny. From this perspective, this section examines the main regulatory and institutional frameworks currently in place, emphasizing the European context, but with an eye on converging global initiatives.

The AI Act establishes a risk-based regulatory framework for AI systems. The regulation classifies AI systems according to risk and imposes corresponding obligations, such as conformity assessments, notifications, technical record keeping, and post-market monitoring. However, as discussed in the previous section, criticism persists regarding the lack of specific mechanisms for supervising notified bodies and the limited external monitoring instruments. Studies indicate that the model proposed by the AI Act still lacks operational guidelines on how to guarantee the integrity and independence of auditors responsible for verifying compliance (Mökander et al., 2022). Although the regulatory text outlines technical and administrative obligations for AI providers, it does not define specific parameters for auditing the bodies responsible for certification, which raises doubts about the effectiveness of regulatory control.

The literature also indicates that institutional reforms should follow a sequenced logic. Priorities include: (i) establishing documentation and traceability infrastructures (Ada Lovelace Institute, 2024; ISO/IEC 42001); (ii) strengthening the independence and regulatory capacity of supervisory authorities (Mökander et al., 2022); and (iii) developing participatory and contestability mechanisms that redistribute epistemic power (Vecchione, Levy & Barocas, 2021). This sequence emphasizes that governance capacity must precede technical expansion to prevent capture and ensure meaningful accountability.

In parallel with the European Union's regulatory initiative, international efforts have expanded to define technical and institutional parameters for auditing AI systems. Notable among these efforts is the Ada Lovelace Institute's *Code & Conduct* report (2024), the Standards Council of Canada (SCC) accreditation program, the ISO/IEC 42001:2023 standard, and the initiatives of the International Association of Algorithmic Auditors (IAAA). These initiatives converge to overcome the limitations of the self-regulatory model, proposing to adopt more robust criteria for traceability, documentation, and independent

verification. Although at different stages of maturity, these instruments seek to structure an audit ecosystem based on multiple layers of control and guided by principles of public legitimacy, regulatory interoperability, and procedural transparency.

The following sections examine the main regulatory, institutional, and technical pathways currently available, or under development, for improving AI system audits. Section 2.1 provides a critical analysis of the AI Act and its delegated regulations, emphasizing the structure and limitations of independent audit mechanisms and the performance of notified bodies. Section 2.2 discusses the concept of dynamic auditing, aimed at the continuous monitoring of systems after implementation, and evaluates the technical and institutional requirements necessary for its effectiveness. Finally, Section 2.3 explores the construction of collaborative ecosystems and the proposal of meta-audit structures, drawing on international experiences involving public participation, independent scientific research, and socio-technical accountability mechanisms.

Regulatory Frameworks and Compliance Mechanisms

The AI Act represents the most comprehensive legislative initiative to regulate AI within a democratic legal framework. Its structure adopts a risk-based approach, classifying AI systems into four categories: prohibited, high-risk, limited-risk, and minimal-risk. Systems considered high-risk include applications in sensitive sectors such as critical infrastructure, employment relations, administration of justice, and law enforcement, including policing, migration control, and border management, as well as applications related to access to essential services and benefits, including health (see Arts. 6 and 8 and Annex III of Regulation (EU) 2024/1689).

For systems classified as high-risk, the AI Act requires suppliers to undergo one of three conformity assessment procedures: (i) self-assessment based on criteria defined in the Regulation and harmonized standards; (ii) evaluation of technical documentation with the participation of a notified body; or (iii) full certification by a notified body (Art. 43). The Regulation also sets out strict requirements for the designation, independence, and accountability of notified bodies (Arts. 31–34), including periodic audits and ongoing supervision by the notifying authorities of Member States. However, as Mökander et al. (2022) observe, gaps remain regarding independent structures capable of exercising continuous oversight over authorized auditors,

particularly concerning methodological transparency and systematic public accountability. The absence of meta-audit mechanisms, social participation, or cross-audits conducted by external entities limits the capacity of the European model to prevent information asymmetries and institutional capture in socially sensitive contexts.

Within the broader European Union regulatory ecosystem, the European Commission adopted Delegated Regulation (EU) 2024/1773, which establishes minimum criteria for the activities of notified bodies, including periodic audits, impartiality requirements, and documentation obligations. Although this Delegated Regulation represents progress in consolidating operational and internal governance parameters, its mechanisms remain focused on indirect state supervision and do not provide for independent external audits or public meta-audit structures. Consequently, transparency and systematic public scrutiny of algorithmic certification processes remain limited (European Commission, 2024). The AI Act (Regulation (EU) 2024/1689), adopted on 13 June 2024, still lacks operational mechanisms to ensure the independence and transparency of notified bodies, particularly regarding continuous supervision, methodological disclosure, and external accountability mechanisms (Mökander et al., 2022; European Union, 2024).While the European framework remains the most advanced legislative experiment in AI regulation, a broader transnational field is emerging to define technical, ethical, and institutional standards for algorithmic audits. These international initiatives seek to fill the gaps left by national and supranational legislation, providing more detailed assessment parameters compatible with legal and organizational contexts. A key example is ISO/IEC 42001:2023, published in December 2023, which defines requirements for AI management systems focused on traceability, continuous improvement, and the integration of governance principles across the algorithmic life cycle.

In the field of ethical standardization, IEEE has promoted guidelines aimed at incorporating human values into the development and application of autonomous and intelligent systems. The document *Ethically Aligned Design* (2023) establishes a normative vision oriented towards human well-being, arguing that AI design should prioritize human dignity, agency, and fundamental rights. This approach provides the foundation for developing the IEEE P7000 series of standards, which seeks to operationalize ethical principles through technical requirements, including traceability, interpretability, and explainability of systems. Still evolving, these standards aim to expand the

scope of algorithmic audits by integrating social and moral requirements into technical evaluation processes.

This perspective is corroborated by the *Code & Conduct* report (Ada Lovelace Institute, 2024), which proposes that AI audits be understood as sociotechnical and relational processes, involving not only technical verifications, but also ethical commitments, transparency obligations, and effective public accountability mechanisms.

Among the certification models under development, the program launched in 2024 by the SCC stands out, establishing formal criteria for the accreditation of entities responsible for auditing AI systems. The proposal seeks to align local regulatory governance with international interoperability standards to ensure greater institutional reliability. This initiative is complemented by the work of the Canadian Standards Association (CSA), which proposes peer auditing and cross-certification mechanisms between independent institutions to reduce the risks associated with self-regulation among private companies in the sector. In both cases, the importance of subjecting certifying entities to external and regular oversight is recognized as a condition for the legitimacy and effectiveness of algorithmic evaluation processes.

Building on the consolidation of regulatory and technical initiatives such as those led by the Ada Lovelace Institute, ISO, and IEEE, the International Association of Algorithmic Auditors (IAAA) has assumed a particularly relevant role in shaping the ethical and operational guidelines of algorithmic auditing at the global level. Adopting a sociotechnical approach, the association emphasizes that AI systems are not merely technical artifacts, but rather the outcome of institutional, social, and political arrangements. From this perspective, the IAAA Code of Conduct (2024) proposes that auditors act with independence, fairness, and responsibility, explicitly considering the distributive effects of automated decisions and the risks of discrimination, exclusion, and opacity. Complementarily, the IAAA AI Auditing Definitions (2024) underline that effective audits must integrate multiple dimensions of analysis—including transparency, accountability, interpretability, and epistemic justice—while ensuring the active participation of communities potentially affected by algorithmic systems. In doing so, the IAAA advances a comprehensive vision of auditing as a process that transcends the verification of codes and data, demanding the scrutiny of institutional and organizational structures that sustain algorithmic practices, and reinforcing the normative commitment to integrity, contestability, and democratic governance.

Algorithmic audits have been widely promoted as a technical solution to the risks of artificial intelligence, but they are insufficient to ensure effective accountability. As proposed by the AI Now Institute (2023), it is necessary to go beyond the limits of the traditional audit model and build robust institutional structures, with enforcement power and the ability to impose corrective measures, suspensions, or sanctions when systems cause harm. Meaningful accountability requires technical verification of compliance and the insertion of audits into broader regulatory ecosystems, involving the active participation of civil society and state mechanisms of public control.

This criticism converges with the proposal for epistemic justice defended by Costanza-Chock et al. (2022), according to which audits should incorporate participatory and community approaches. For the authors, the groups most affected by technologies need an active voice in defining the audit criteria, methods, and objectives, as a condition for their social legitimacy and contextual sensitivity. Both perspectives point to the need to transform audits into public accountability instruments, not just technical processes conducted under corporate logic.

In this sense, sociotechnical audits are based on continuous evaluation processes, with the involvement of multiple actors and attention to the concrete effects of systems on specific groups and contexts. Institutional structures capable of ensuring independence, adaptability, and openness to public scrutiny are essential for such audits to be viable. Although still incipient in most national legislations, these practices point toward a necessary transformation of algorithmic governance, one guided by democratic legitimacy and social justice.

Dynamic Auditing and Post-Market Monitoring

While traditional audits focus on discrete and static evaluations, typically conducted during the development phase or immediately prior to system deployment, the advancement of AI in dynamic and adaptive contexts requires continuous verification models. In this scenario, dynamic auditing presents itself as a more appropriate response to new complexities by proposing iterative inspections that remain sensitive to transformations occurring throughout the entire lifecycle of automated systems. As Maughan et al. (2022) argue, compliance should not be treated as a final state, but as a permanent condition, which requires constant monitoring, periodic reassessment, and updated testing in the face of changes in input data, application contexts and the AI -models

themselves, especially in the case of adaptive systems or systems based on *online machine learning*, understood here as algorithms that update their parameters with the arrival of new data, rather than "online learning" in the sense of remote or distance education.

Dynamic auditing can be understood as the inspection that accompanies AI systems over time. It is based on three main ideas. The first is the need to periodically review systems, through continuous assessment cycles and updates with new data. The second is the flexibility of audit methods, which must adapt to different practical contexts and legal requirements. The third is maintaining communication channels between auditors, developers, and affected users, ensuring that the assessment considers diverse perspectives and real user experiences (Maughanet al,. 2022).

Practical examples reinforce the urgency of this approach. In automated systems used by large corporations and public agencies, **neutral adjustments to decision criteria can inadvertently generate significant distortions along sensitive attributes such as race, gender, or geographic origin.** The report by the *Electronic Privacy Information Center* (Fergusson, 2023) shows that, even after declared adjustments, third-party automated systems continued to present structural flaws, compromising equity and making it difficult to hold those involved accountable.

The problem becomes particularly visible in recommendation and personalization systems, where rapid adaptation to user behavior amplifies informational asymmetries. Mittelstadt (2016) argues that such systems can compromise informational diversity and, consequently, the cognitive autonomy of users. By filtering content based on profiles and past behavior, these systems may create self-reinforcing informational environments, often described as "algorithmic bubbles", that hinder exposure to diverse perspectives and constrain pluralistic public debate. Although influential, this notion has also been criticized as overly reductive, since empirical evidence for fully isolated "filter bubbles" is mixed. Still, the underlying concern remains relevant: opaque personalization mechanisms can weaken the conditions for informed democratic participation.

The European regulatory framework for high-risk AI systems requires vendors to establish ongoing processes for monitoring system performance after implementation, as described in the delegated regulations that complement the AI Act. These processes include data collection, risk analysis, independent audits, and systematic documentation. **However, primary responsibility for these activities remains with vendors, which undermines the effectiveness of**

the model in the absence of explicit meta-audit mechanisms or mandatory external oversight (European Union, 2024).

This regulatory framework has been criticized for emphasizing self-regulation, especially in technically complex and socially sensitive sectors. As Mökander et al. (2022) point out, the reliance on self-diagnosis by developers themselves undermines the credibility of the governance model and perpetuates structural asymmetries between the companies responsible for the systems and the regulatory bodies. This dynamic weakens effective institutional control and makes it challenging to detect failures promptly, generating additional risks in applications with high social impact.

Institutional and technical challenges stand out among the main barriers to the effective implementation of dynamic audits. On the one hand, many regulatory authorities lack infrastructure, qualified personnel, and timely access to data and models, making it challenging to exercise autonomous and timely oversight powers. On the other hand, there are structural challenges in the design of AI systems themselves, which are not always designed to be auditable from the outset. Rastogi et al. (2023) demonstrate that, even in technically sophisticated environments involving large language models (LLMs), auditability depends on providing specific tools, contextual organization of failures, and active interaction with human auditors. The study highlights that limitations such as the lack of interpretable logs, standardized documentation, and adequate interfaces for exploring algorithmic behaviors hinder the effectiveness of audits, especially those of a dynamic nature.

Despite recent regulatory advances, continuous audit enforcement remains incipient in most jurisdictions. The Ada Lovelace Institute's *Code & Conduct* report (2024) and initiatives such as the SCC (2024) program outline essential guidelines for the institutional structuring of independent audits. However, both lack mandatory legal mechanisms to ensure their systematic implementation. As the Holistic AI report (2023) warns, the lack of regulatory standardization and mandatory requirements for dynamic audits favors an asymmetric scenario, in which only companies with a greater reputation or resources voluntarily adhere to good practices, compromising the regulatory ecosystem's fairness and effectiveness.

In short, dynamic auditing represents a necessary evolution of the algorithmic verification paradigm, aligned with the principles of adaptive justice, continuous surveillance, and iterative correction. Its consolidation, however, will depend on profound institutional reforms, strengthening oversight capabilities and enforcing technical standards that make auditability a legal obli-

gation rather than an optional choice for AI systems with significant societal impact.

Collaborative Ecosystems and Meta-Audit Infrastructure

The evolution of the algorithmic auditing debate underscores that technical procedures alone are insufficient to ensure legitimacy, effectiveness, and fairness in the evaluation of AI systems. In contexts characterized by epistemic asymmetries, institutional inequality, and technological opacity, audits require broader sociotechnical infrastructures capable of integrating diverse social actors, including civil society organizations, universities, regulatory bodies, and independent technical communities.

In this debate, sociotechnical auditing has been discussed as an alternative approach that emphasizes the need to examine not only the technical performance of AI systems but also the institutional, political, and social conditions under which they operate (Lam et al., 2023).

As argued by Lam et al. (2023), sociotechnical audits seek to reveal how algorithmic decisions are shaped by power dynamics and social relations, which requires an interdisciplinary approach, with the active participation of experts in ethics, law, engineering, and social sciences. Such an approach also poses the challenge of including, in the audit processes, the voices and experiences of the communities most directly affected by automated systems, especially those historically marginalized in decision-making arenas. The research by Rastogi et al. (2023) documents the development of *AdaTest++*, an audit tool that combines the strengths of human auditors and generative language models. The study demonstrates that human–LLM collaboration can enhance audit effectiveness by enabling the generation of more diverse tests, the contextual organization of failures, and the iterative inspection of algorithmic behavior.

Among the most advanced proposals to institutionalize this approach, the public or hybrid meta-audit model, advocated by the Ada Lovelace Institute (2024), stands out. The concept of meta-audit refers to the creation of independent mechanisms to monitor the auditors themselves, to ensure that verification processes are conducted ethically, transparently, and technically valid. The *Code &Conduct* report (Ada Lovelace Institute 2024) recommends that AI audit regimes establish independent and transparent structures subject to public scrutiny, involving external actors such as civil society organizations, independent experts, and public institutions to ensure verifiable ethical and technical accountability.

The European Union, through the AI Act, establishes that the conformity assessment function falls to notified bodies, independent conformity assessment organizations designated by Member States and authorized to evaluate whether high-risk AI systems meet the requirements of the Regulation. These entities may, under their responsibility and with the supplier's consent, subcontract parts of the process or use subsidiaries, as provided for in Article 33 (European Union, 2024). This subcontracting, however, is subject to strict criteria and does not constitute an institutionalized form of cooperation with external independent or public entities. On the other hand, Canada has advanced toward a more distributed governance model through the SCC accreditation program, which imposes criteria of accountability, independence, and transparency on certifying entities (SCC, 2024) and proposes the creation of technical public audit centers. In turn, the CSA has promoted peer audit projects in partnership with universities and applied research centers (CSA, 2024).

Although there is no comprehensive federal regulation on artificial intelligence in the United States, several decentralized initiatives have been implemented. The National Institute of Standards and Technology (NIST) developed the *AI -Risk Management Framework* (AI RMF 1.0), a voluntary guide released in January 2023, which aims to help organizations manage risks associated with AI by promoting trustworthy and responsible systems. This framework is structured into four main functions: Govern, Map, Measure, and Manage (NIST, 2023). In parallel, the Federal Trade Commission (FTC) has taken steps to ensure the responsible use of AI. Per the *Office of Management and Budget Memorandum M-24-10*, the FTC has developed a compliance plan emphasizing transparency, accountability, and a focus on public benefit in using AI tools (FTC, 2024).

Among the emerging actors, the IAAA stands out, seeking to consolidate a global ethical-normative repertoire for AI audits, promote the training of auditors, and create an international public certification registry (IAAA, 2024). Holistic AI has also contributed technical proposals on audits of large language models, emphasizing good operational practices, the documentation of procedures, and the responsible use of metrics in complex environments (Holistic AI, 2023).

Establishing a global meta-audit infrastructure is therefore both a technical and a political imperative. It requires strengthening the institutional capacity of states and independent organizations, creating transparent and interoperable protocols, and institutionalizing mechanisms for social participation in algorithmic governance. As Vecchione, Levy, and Barocas (2021) argue,

algorithmic audits should be evaluated not only for their technical robustness but also for their capacity to empower those affected by automated systems to contest their outcomes, an endeavor that presupposes structures open to public criticism and continuous review. Overcoming the closed self-regulation of private audits is essential to building and maintaining public trust, ensuring that AI systems operate in accordance with democratic values and fundamental rights.

Limitations

This study is based on a normative and documentary analysis and does not include primary empirical data. The literature reviewed highlights structural constraints that limit deeper empirical investigation, such as the lack of standardized documentation and comparable audit records (Costanza-Chock et al., 2022), the technical opacity of systems subject to audit (Mittelstadt, 2016), and restricted public access to models, logs, and evidence required for detailed analysis (Fergusson, 2023). In line with Schiff, Kelley, and Camacho (2024), this study acknowledges that the absence of interviews or fieldwork limits the examination of real audit practices and potential dynamics of institutional capture. These constraints reflect broader challenges in the emerging field of AI audit research and should be addressed in future studies.

Conclusion

This study addressed the central question, "Should AI auditors be audited?" Rather than a merely technical or regulatory issue, this study argues that the question reflects deeper concerns about the power structures shaping contemporary algorithmic governance. The analysis has shown that, although conceived as accountability instruments, AI audits often operate in ecosystems marked by epistemic asymmetries, conflicts of interest, and opaque self-regulation.

The discussion identified structural flaws across multiple dimensions: technical, when audits privilege quantitative metrics and neglect social impacts; institutional, when auditors depend economically on those being audited; and regulatory, due to weak external oversight mechanisms, including the European model centered on notified bodies. Without ongoing public

supervision, reliance on self-assessment and certification risks undermining the legitimacy and effectiveness of audits. These findings should also be interpreted considering the methodological constraints discussed earlier, particularly the limited availability of standardized documentation and the opacity of systems subject to audit, which restrict the empirical visibility of real auditing practices.

To confront these limitations, the study proposes the creation of permanent public or hybrid meta-audit structures grounded in transparency, integrity, and institutional independence. Such a reconfiguration entails inter-institutional cooperation among regulatory agencies, universities, civil society organizations, and technical communities, fostering democratic forms of supervision and contestation.

Ultimately, auditing should be understood not merely as a technical verification mechanism but as a political and epistemic practice that determines who defines standards and interprets social effects. Ensuring independent audits of auditors is therefore both a demand of epistemic justice and a democratic safeguard. Meta-audits, by institutionalizing independent and participatory oversight, are essential to protect fundamental rights and reinforce the democratic legitimacy of algorithmic governance.

References

Ada Lovelace Institute. *Code & Conduct: How to Create Third-Party Auditing Regimes for AI*. Londres, 2024. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2024/06/Ada-Lovelace-Institute-Code-and-conduct-FINAL-1906.pdf.

Ada Lovelace Institute; AI Now Institute; Open Government Partnership. *Algorithmic accountability for the public sector: learning from the first wave of policy implementation*. London: Ada Lovelace Institute, 2021. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.

AI Now Institute. *Algorithmic Accountability: Moving Beyond Audits*. 2023. Available at: https://ainowinstitute.org/publications/algorithmic-accountability.

Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. ProPublica*. New York, 2016a.

Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. ProPublica*. New York, 2016b. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Bandy, Jack. *Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits*. arXiv, 2021. Available at: https://arxiv.org/abs/2102.04256.

Bouwmeester, M. *System Failure in the Digital Welfare State: Exploring Parliamentary and Judicial Control in the Dutch Childcare Benefits Scandal*. University of Groningen, 2023. Available at: https://research.rug.nl/en/publications/system-failure-in-the-digital-welfare-state-exploring-parliamentar.

Costanza-Chock, S.; Raji, I. D.; Buolamwini, J. *Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem*. In: ACM Conference on Fairness, Accountability and Transparency – FAccT '22, 2022, New York. *Proceedings*. New York: Association for Computing Machinery (ACM), 2022. p. 1571–1583. Available at: http://dx.doi.org/10.1145/3531146.3533213.

Canadian Standards Association (CSA). *Advancing responsible AI: SCC launches accreditation program for AI management systems*. 2025. Available at: https://www.scc.ca/en/news-events/news/launch-ai-governance-accreditation.

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review. *The Academy of Management Review,* 14(1), 57–74. Available: https://doi.org/10.2307/258191.

European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. Official Journal of the European Union, L 168, 12.7.2024, p. 1–138. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.

European Commission. *Commission Delegated Regulation (EU) 2024/1773 of 1 March 2024 supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council as regards the request for exemption from the obligation to create an advertisement repository by providers of very large online platforms*. Official Journal of the European Union, L 177, 26.6.2024. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401773.

Federal Trade Commission (FTC). *Artificial intelligence compliance plan*. Oct. 2024. Available at: https://www.ftc.gov/ai.

Fergusson, G. *Outsourced automated: how AI companies have taken over government decision-making*. Electronic Privacy Information Center (EPIC), 2023. Available at: https://epic.org/wp-content/uploads/2023/09/FINAL-EPIC-Outsourced-Automated-Report-w-Appendix-Updated-9.26.23.pdf.

Holistic AI. *LLM auditing guide: what it is, why it's necessary, and how to execute it.* Oct. 11, 2023. Available at: https://www.holisticai.com/papers/llm-auditing-guide.

Institute of Electrical and Electronics Engineers (IEEE). *Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems.* New York: IEEE, 2023. Available at: https://sagroups.ieee.org/global-initiative/wp-content/uploads/sites/542/2023/01/ead1e.pdf.

International Association of Algorithmic Auditors. *IAAA AI auditing definitions.* [S.l.]: IAAA, Oct. 2024. Available at: https://iaaa-algorithmicauditors.org/wp-content/uploads/2024/10/IAAA-AI-Auditing-Definitions-1.pdf.

International Association of Algorithmic Auditors. *IAAA code of conduct.* [S.l.]: IAAA, Oct. 2024. Available at: https://iaaa-algorithmicauditors.org/wp-content/uploads/2024/10/IAAA_Code_of_Conduct-2.pdf.

ISO/IEC. *ISO/IEC 42001:2023 – Artificial intelligence — management system*. International Organization for Standardization, 2023. Available at: https://www.iso.org/standard/81230.html.

Lam, M. S. et al. *Sociotechnical audits: broadening the algorithm auditing lens to investigate targeted advertising*. *Proceedings of the ACM on Human-Computer Interaction*, v. 7, n. CSCW2, Article 360, p. 1–37, Oct. 2023. DOI: https://doi.org/10.1145/3610209.

Maughan, K.; Ngong, I. C.; Near, J. P. *Prediction sensitivity: continual audit of counterfactual fairness in deployed classifiers.* 2022. Available at: https://arxiv.org/pdf/2202.04504.

Mittelstadt, B. *Auditing for transparency in content personalization systems. International Journal of Communication*, 2016. Available: https://www.researchgate.net/publication/309136069.

Mökander, J.; Axente, M.; Casolari, F.; Floridi, L. *Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. Minds & Machines*, v. 32, p. 241–268, 2022. Available: https://doi.org/10.1007/s11023-021-09577-4.

National Institute of Standards and Technology (NIST). *AI risk management framework (AI RMF 1.0)*. Jan. 2023. Available at: https://www.nist.gov/itl/ai-risk-management-framework.

Rastogi, C.; Sugimoto, C. R.; Maddox, T. et al. *Supporting human-AI collaboration in auditing LLMs with LLMs.* 2023. Available at: https://arxiv.org/pdf/2304.09991.

Standards Council of Canada (SCC). *AI governance accreditation program*. 2024. Available at: https://www.scc.ca/en/news-events/news/launch-ai-governance-accreditation.

Schiff, D. S.; Kelley, S.; Camacho Ibáñez, J. *The emergence of artificial intelligence ethics auditing. Big Data & Society*, 2024. Available at: https://journals.sagepub.com/doi/10.1177/20539517241299732.

Stigler, G. J. *The theory of economic regulation. The Bell Journal of Economics and Management Science*, v. 2, n. 1, p. 3–21, 1971. Available: https://doi.org/10.2307/3003160.

Stilgoe, J. *We need a Weizenbaum test for AI. Science*, v. 381, n. 6658, p. 770–771, 2023. Available: https://www.science.org/doi/10.1126/science.adk0176.

Treleaven, P.; Galpin, V.; Zardousti, P. *Algorithmic auditing: AI accountability in public sector applications.* SSRN, 2021. Available at: https://ssrn.com/abstract=3889612.

Vecchione, B.; Levy, K.; Barocas, S. *Algorithmic auditing and social justice: lessons from the history of audit studies.* In: *Equity and Access in Algorithms, Mechanisms, and Optimization – EAAMO '21*, In Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). New York: ACM, 2021. p. 1–9. Available: http://dx.doi.org/10.1145/3465416.3483294.

Wörsdörfer, M. *The E.U.'s Artificial Intelligence Act: an ordoliberal assessment.* Forthcoming in: *AI and Ethics*, 2023. Available at: https://ssrn.com/abstract=4544276 or http://dx.doi.org/10.2139/ssrn.4544276.