A Framework for Assurance Audits of Algorithmic Systems

KHOA LAM*, BABL AI Inc., USA
BENJAMIN LANGE*, BABL AI Inc., USA and Ludwig Maximilians University, Germany
BORHANE BLILI-HAMELIN*, BABL AI Inc., USA and AI Risk and Vulnerability Alliance, USA
JOVANA DAVIDOVIC, BABL AI Inc., USA and University of Iowa, USA
SHEA BROWN, BABL AI Inc., USA and University of Iowa, USA
ALI HASAN, BABL AI Inc., USA and University of Iowa, USA

An increasing number of regulations propose the notion of 'AI audits' as an enforcement mechanism for achieving transparency and accountability for artificial intelligence (AI) systems. Despite some converging norms around various forms of AI auditing, auditing for the purpose of compliance and assurance currently have little to no agreed upon practices, procedures, taxonomies, and standards. We propose the 'criterion audit' as an operationalizable compliance and assurance external audit framework. We model elements of this approach after financial auditing practices, and argue that AI audits should similarly provide assurance to their stakeholders about AI organizations' ability to govern their algorithms in ways that mitigate harms and uphold human values. We discuss the necessary conditions for the criterion audit, and provide a procedural blueprint for performing an audit engagement in practice. We illustrate how this framework can be adapted to current regulations by deriving the criteria on which 'bias audits' for hiring algorithms can be performed, as required by the recently effective New York City Local Law 144 of 2021. We conclude by offering critical discussion on the benefits, inherent limitations, and implementation challenges of applying practices of the more mature financial auditing industry to AI auditing where robust guardrails against quality assurance issues are only starting to emerge. Our discussion as informed by experiences in performing these audits in practice highlights the critical role that an audit ecosystem plays in ensuring the effectiveness of such methodology.

CCS Concepts: • **Social and professional topics → Governmental regulations**; **Technology audits**.

Additional Key Words and Phrases: AI auditing, AI regulation, algorithm audits, bias audits, AI bias, accountability, disparate impact testing

1 INTRODUCTION

Auditing has been proposed in a variety of laws and regulations [17, 21, 22, 24, 40, 49, 71],[1] standardized frameworks [45, 58], and guidelines for industry best practices [5, 73] as a mechanism to identify and mitigate risks of harm in artificial intelligence (AI)[2] and to build public trust and promote accountability for AI system developers and deployers. Most notably, New York City (NYC) enacted the Local Law 144 of 2021 ('NYC bias audit law' hereafter) in which *bias audits*—defined as independently conducted impartial evaluations—are required for all automated tools used in hiring and promotion [71]. Auditing for the purpose of compliance and assurance with normative requirements currently lacks defined norms and standardized practices, despite notable emerging efforts to perform audits as many types of engagement, including adversarial pressure testing [3, 64, 76], quantitative technical assessments [13, 16, 20, 36, 78], and qualitative assessments of risks or impacts [10, 18, 29], among others.

In this paper, we propose a criterion audit framework for the external compliance and assurance audits of algorithmic systems. This approach is inspired by how financial audits are used to provide assurance that financial statements are

*Authors contributed equally to this research.

[1] Mökander et al. [43] argue that *conformity assessments* as an enforcement mechanism in the EU AI Act should be interpreted as auditing, despite the term 'audits' not being referred to explicitly.

[2] In this paper, we use 'AI,' 'algorithm,' 'AI system,' and 'algorithmic system' interchangeably. Similarly, 'AI audit' and 'algorithm audit' both convey the same meaning.

presented accurately and in conformity with generally accepted accounting standards (GAAS) [28]. Our aim is to show *how* and—in part—*to what extent* the methodology and practices of such a mature industry can be applied to AI auditing as an emerging industry absent of agreed upon standards and official certification bodies. While the idea to follow the financial auditing practices—at least in part—is in itself not new [26, 42, 54], our paper offers an actionable approach for performing external compliance and assurance audits which is missing in the literature. Furthermore, as we have also used this framework to perform audits for the NYC bias audit law, our critical discussion draws on practitioner perspectives from operationalizing these external audits in the field.

Our paper proceeds as follows. Section 2 provides relevant background on the AI auditing landscape, including the operational gaps in the current regulations, and a short survey on taxonomy. We then introduce the criterion audit framework in detail. Section 3 outlines the (1) primary objectives of the audit framework, (2) its key elements, (3) the approaches to auditing supported by this framework, (4) the procedures for executing this audit, and lastly, (5) the responsibilities of an auditor tasked with its performance. In Section 4, we demonstrate a specific use case of the framework by adapting it to NYC Local Law 144 of 2021. Section 5 critically examines the benefits, limitations, and implementation challenges posed by this framework. Finally, section 6 concludes our discussion.

2 BACKGROUND

2.1 Motivation

In recent years, policymakers have introduced various forms of formal evaluation of algorithmic systems as a prominent mechanism to bring about transparency and accountability, with some requiring that these algorithmic systems undergo ‘audits’ [17, 21, 22, 24, 40, 49, 71]. However, most proposed policies have offered little guidance on audit quality assurance issues, such as independence rules, engagement performance matters, auditor qualification, and quality control procedures [38]. Among these, legislation has focused predominantly on independence. While some regulations, such as NYC bias audit law and EU Digital Services Act, have set rules around auditor eligibility, others, such as The Algorithmic Accountability Act of 2020 and the EU AI Act, offer no guidelines on independence at all.

Standardized procedures remain largely unaddressed by regulations. Although there are emerging efforts to provide guidance for the execution of audits [12, 23], there is no consensus regarding how audits ought to be performed. Auditors are thus left to their own devices to operationalize audit engagements [18]. This lack of clear guidance poses several challenges in the current algorithm auditing industry. First, there is a significant risk of inconsistency in audit quality, due to the discrepancies in audit engagement performance. This issue can eventually lead to failure to achieve the aims of the regulation for which audits were performed. Second, without standard practices, auditors themselves may face issues of liability for risks concerning false assurance [27].

Against this landscape, our discussion aims to (1) provide a framework that answers central questions about how auditing for the purpose of compliance or assurance can be conceptualized, structured, and applied in practice, and (2) critically consider its key benefits and limitations.[3]

[3] See [27, p. 294–296] for key questions that the audit regime has to answer.

2.2 Taxonomy

Given the flexible nature of audits in other contexts, the term 'audit' has been borrowed to describe many disparate forms of evaluation of algorithmic tools, products, and systems [18, 27]. Audit engagements can range from pressure-testing efforts by journalists and civil society without direct access to the audited system [3, 36, 64, 76],[4] to pre-deployment evaluations conducted by teams internal to an organization [53], to 'collaborative audits' conducted by an outside team without safeguards against conflicts of interest [78], to audits conducted by outside parties with access to a system under robust safeguards against conflicts of interest [54]. Here, we focus specifically on 'external audits' aimed at providing assurance of compliance with requirements set forth in legislation or other standardized frameworks, which are sometimes also referred to as 'compliance audits' [1, 44].

Internal audits—such as the model evaluations required by the US Federal Reserve's Supervisory Letter (SR) 11-7 [9]—are forms of self-assessment: evaluation conducted by the audited party or by outside contractants without robust safeguards against conflicts of interest [2, 54]. By contrast, external audits are evaluations conducted by independent parties outside of the audited entity, such as the public or regulators. In the US, relevant models of external audits from other fields include outside evaluations conducted by certified independent third parties—such as financial audits that conform to the Sarbanes–Oxley Act of 2002 [55] and Leadership in Energy and Environmental Design (LEED) certification by the US Green Building Council—and evaluations conducted by a government agency, such as FDA inspections for food and drugs [2, 54].[5] There is a growing consensus that, with internal audits, conflicts of interest between the auditor and the auditee may undermine their purpose of providing the assurance of trustworthiness or compliance to outside parties [2, 22, 27, 54, 62, 71]. Potential measures for ensuring independence and preventing conflicts of interest in external audits include: rules against cross-selling of non-audit services, legal liability of auditors for false provision of assurance, standardization of audit criteria, professionalization of audits, creation of 'auditing intermediaries' [75],[6] government selection of auditors, among others [27, 54].

Evaluations performed by external parties without adequate access to the audited system—referred to as 'critical third party audits' by Metcalf et al. [41]—can play an important role in uncovering and drawing public attention to existing flaws of deployed AI systems. They aim to inspire accountability by way of forcing organizations to 'fix' their AI systems (through technological or governance means) or face public scrutiny. However, we follow Raji et al. [54, p. 558] in assuming that this form of third party oversight without adequate access "[is] not conventionally considered audits per se." By contrast, our proposed audits require formal auditor access into the audited system. Furthermore, a more fundamental distinction concerns its public role, where criterion audits, similar to financial audits, function to provide 'comfort' by way of assurance and certification.[7]

Finally, we also consider audits distinct from risk or impact assessments [63]. First, proposals for algorithmic impact or risk assessments typically focus on internal assessments [54, 60]. Second, as we examine in this paper, effective external audits require a determination of an outcome (e.g., whether an audit passes or fails) such that it can enable stakeholders of the audit to act accordingly. By contrast, risk or impact assessments are better understood as having open-ended outputs, such as a prioritized and normatively justified list of risks or impacts [29, 60].

[4] We note that while journalists do not typically refer to their works as 'audits' but as 'analyses' or 'investigations,' these efforts are nonetheless widely cited as exemplary audits in the reviewed literature.

[5] On the long history of independent audits in finance, see [27, 50, 51].

[6] An 'auditing intermediary' refers to a proposed specialized audit entity whose responsibility is to ensure that accurate and verified data from large online platforms is received by regulators when shared across jurisdictions.

[7] See [51, p. 126] for discussion on public function of financial audits. In AI auditing, tensions also arise from disagreements about which role audits ought to play [42].

3 THE CRITERION AUDIT

3.1 Objectives

The primary objective of financial auditing is to provide assurance to stakeholders (e.g., investors, lenders, and regulators) regarding the reliability of an organization's financial information for informed decision-making. Financial auditing assumes a vital role in upholding public trust in the financial system and ensuring economic stability by holding organizations accountable for their financial reporting and operations. External compliance and assurance algorithm audits should similarly aim to offer stakeholders reassurance that algorithmic systems are designed, built, deployed, and governed in a responsible and transparent manner. While AI systems do not bear foundational responsibility for upholding the integrity of the financial market, we argue that the opaque, rapidly transformative, and increasingly pervasive nature of AI systems in everyday life demands an assurance that algorithmic systems function in ways that mitigate harms and uphold human values.[8] AI audits can thus be employed as a mechanism to provide this desired reassurance and trust between stakeholders of AI by way of systematization and transparency [42].

If transparency is a key mechanism for AI audits to foster trust, current practices in the algorithm auditing industry often fall short of achieving such ideal. Most auditing efforts do not publish outcomes or procedures to the public, often due to clients' confidentiality agreements [18]. Many proposed regulations of AI auditing also do not require disclosure of an outcome. This lack of transparency can gradually degrade accountability by rendering auditing toothless, as audited organizations are not incentivized through either public pressure or regulatory enforcement to create meaningful change. Furthermore, due to procedural non-transparency, audit process effectiveness is not measurable, perpetuating what Power [50] refers to as the 'regress of mistrust,' where accountability and trust are shifted away from the audited organizations and onto auditors themselves, who are in turn subjected to more audits and evaluations. Our proposed audit framework aims to provide this critical and meaningful layer of transparency by way of public disclosure requirements (see 3.2 and 3.4.4). While public disclosure is not a comprehensive solution to achieve accountability,[9] our proposal aims to be sufficiently consequential in a way that it enables audit stakeholders (e.g., regulatory enforcers, audit report readers, impacted communities, the public, and developers/deployers themselves) to take actions towards advancing accountability for AI companies around the design, development, deployment, and governance of their algorithms.

3.2 Definition

We define a *'criterion audit'* as:

> **Criterion Audit:** A criteria-based independent external evaluation E of an algorithmic system S conducted by an auditor A to determine whether the given system S meets the requirements set by a normative framework.

Audit criteria are the set of verifiable or observable conditions that must be jointly satisfied for an algorithmic system S to count as compliant with a given standard or law. These conditions must enable auditors to form an unambiguous opinion about whether a given criterion is satisfied based on evidence they can obtain within the operational context of the audit. For example, as we show in later sections, in the case of NYC bias audit law, these may include the relevant criteria for determining how a given system is assessed for disparate impact. By 'independent third-party evaluator,'

[8] There is convergence on the importance of external audits for providing assurance to outside parties, such as to governments, the public, or other organizations [2, 27, 44, 54, 62].

[9] On the limits of extensive public disclosure in financial audits, see [50, p. 20].

we refer to an auditor that is independent of the customer-supplier relationship and is free of any conflict of interest with respect to the client that is being audited. On this matter, we understand 'independence' as either: (1) lacking any contractual relationship with the auditee, or (2) involving rigorous safeguards against conflicts of interest, in cases where fees related to the audits themselves are paid by the auditee.[10]

Four features about the criterion audit are worth highlighting:

(1) **The audit must be conducted against a set of standardized and publicly accessible criteria**, to provide procedural transparency to the audit. By showing what criteria the algorithmic system was evaluated against, the results of the audit are thus contextualized and more interpretable by readers of the audit report.[11]

(2) **The audit's objective must be to measure compliance or providing assurance against a normative framework**, such as a regulation. This precise and narrow audit scope would allow for measurement and comparison of the audit's performance [18, 42, 54]. Moreover, it also provides the basis for which audit criteria are constructed.

(3) **Auditors must be trained and accredited in a standardized manner** [18, 54]. They must also be held to high professional standards of quality assurance and ethics—e.g., in the form of a Code of Ethics or a standard of audit quality assurance. While audit standards are nascent, training courses and programs have started to appear [25, 32, 65], aiming to provide such standardized certification for auditors.

(4) **Results of the audit must be publicly disclosed, at least in some restricted form**, to meet both the public's need for standard reporting and audited organizations' concerns regarding security and intellectual property [54]. Furthermore, the content of public disclosure can be stipulated by the legislation, and supplemented further to facilitate contextual interpretation of the results. The degree of public disclosure can vary, depending on contextual details. However, at a minimum, it should contain key information that is necessary for readers of the audit report to understand compliance with a given piece of legislation.

3.3 Auditing approaches

A number of auditing approaches can be adapted to be compatible with the features of this audit framework. We introduce two paradigm approaches below:[12]

3.3.1 Direct assessment. The auditor directly performs the procedures on the algorithmic systems, as scoped by the audit. In this approach, the auditor is given—for technical testing—full, or mediated, access to the model to conduct technical testing, or—for a risk assessment—access to relevant stakeholders of the model, such as its impacted users. When the assessments are performed as part of a regulation's requirements, the criterion audit framework also demands (1) that these assessments be performed in accordance with a transparent set of criteria by qualified audit professionals, and (2) that their results be partially disclosed publicly.

3.3.2 Indirect verification. Alternatively, the auditee can perform the assessments themselves, and subsequently submit a body of evidence to prove the performance and quality of these auditable procedures. This requires the auditors to evaluate the procedures, policies, and institutional structures performed and established by the auditee against a set of criteria which examines their quality. This process can act as a complement to the work of an internal audit team

[10] Such is the norm in financial audits, see also [59, 70] for standards on independence for financial auditors.

[11] For non-financial assurance engagements, the International Standards on Assurance Engagements (ISAE) 3000 [66] specifies the characteristics of suitable audit criteria, which includes relevance, completeness, reliability, neutrality, and understandability.

[12] These approaches are based loosely on 'direct engagements' and 'attestation engagements' in financial audits.

who audits the algorithmic system along its development lifecycle,[13] and produces critical documents and artifacts—such as model or ethical risk analysis results, system cards, and technical testing reports [53]. In this way, efficiency is afforded between internal and external auditing functions where the latter reviews and evaluates documentation provided by the former for the specific objective of demonstrating compliance or assurance. Moreover, the criterion audit framework can apply if (1) the evaluation process is conducted by qualified practitioners, and (2) some results of the audit are subsequently publicly disclosed.

There are inherent and practical trade-offs between these approaches. Direct assessment may provide better safeguards against accountability concerns such as rubber stamping, whereas indirect verification may find efficiency in large-scale audit engagements for auditees with internal audit teams. In practice, auditors may utilize one single approach, leverage both in a hybrid manner, or determine the suitable auditing approach based on an analysis prior to performing audit procedures.[14]

In this paper, we propose procedures (in Section 3.4), and auditor responsibilities (in Section 3.5) based on the indirect verification approach. In addition, the criteria designed for NYC bias audit law (in Section 4.1) were also developed assuming the verification approach of auditing.

3.4 Audit procedures

We propose the following standard process to perform a criterion audit:

3.4.1 Target scoping. The auditor scopes the audited algorithmic system to obtain a foundational understanding of its technical and sociotechnical components, in addition to any process deemed relevant to the normative framework, such as whether the system has undergone bias testing. This is intended to provide auditors with a contextual knowledge of the algorithm before evaluating it against a set of criteria. Scoping can be done, for example, in the form of a targeted questionnaire or through interviews with personnel of the audited organization.

3.4.2 Documentation submission. The auditee submits documentary evidence towards satisfaction of the audit criteria, and the auditor reviews the documentation and determines whether the criteria has been preliminarily met, pending verification of evidence (see below). During the review process, if the auditor requires more evidence, they may ask for additional documentation or request interviews with the auditee's internal or external stakeholders, such as employees or third-parties, to facilitate evaluation. The auditor reaches an initial opinion about whether the criteria has been satisfied by the evidence provided by the auditee. Auditors therefore need to have the knowledge and expertise to make qualified critical judgments based on evidence of whether a given criterion has been sufficiently met.

3.4.3 Evidence verification. The auditor determines the truthfulness of the evidence presented by the auditee. Methods for verification, such as examining official communication logs, or observing the re-performance of a computation in an interview, can vary based on the nature and importance of the evidence. For example, the auditee may show the computation of bias testing to verify the results shown as evidence submitted previously. At the end, the auditors reach a final opinion about the satisfaction of the criteria, which takes into account the veracity of the evidence.[15] This mirrors the standard procedures for obtaining evidence in financial audits [52] and builds upon proposals calling for

[13] See also [57] for discussion on The Three Lines of Defense (3LoD) model as a proposed mechanism for risk mitigation in the AI context.

[14] In the Delegated Regulation on auditing for the Digital Services Act [22], auditors are required to conduct an audit risk analysis to select the precise audit methodologies.

[15] In practice, some information required to verify evidence may have already been provided as part of the documentation submission (e.g., communication logs, database queries, event recordings), in which case, the auditor can now determine whether the quality of such evidence is sufficient.

adaptation of claim-based assurance frameworks to the AI context [11, 30]. As auditors also make judgment calls, they need to adhere to strict and rigorous standards in performing the verification.

3.4.4 Publication of the audit report. The auditor drafts and publishes the audit report regarding the audited algorithmic system. Results should be disclosed in a standardized format, and the report at the least needs to explicitly and publicly show (1) whether each criterion has been met, (2) the final outcome of the audit, and (3) a formal opinion of the auditor. Before publication, the auditee may be allowed to review the full report for strictly factual errors or any omissions that may have materially affected the report.

3.4.5 Certification. The audited algorithmic system receives a certification indicating the outcome of the audit—e.g., whether the algorithmic system has passed (or failed) with respect to the targeted regulation against which it was evaluated.

3.5 Auditor responsibility

In financial auditing, the outcome of an audit requires an auditor forming an opinion on the financial statements based on having obtained sufficient appropriate audit evidence about whether these statements are free from material misstatement [67]. Achieving a high-quality audit requires auditors to (1) exhibit professional values and ethics, (2) have sufficient knowledge and skills in their subject matter, and (3) apply rigorous and appropriate audit process and quality control procedures.

An auditor for the criterion audit similarly bears the responsibility of forming an opinion based on the audit evidence provided by the auditee to evaluate the level of appropriateness, sufficiency, and material misstatement for the submitted body of evidence.[16] Determination of appropriateness, sufficiency, and materiality can take many forms, such as conducting statistical testing (e.g., statistical significance, power analysis), evaluating the appropriateness and reasonableness of a normative justification for important choice-points (see 4.2.1), or determining whether an auditee attempts to deceive the auditors or game the audit process. Such judgment calls and decision-making should therefore be made by auditors having not only (1) expertise in specific subject matter (such as technical acumen for technical audits, or expertise in normative ethics or sociology for sociotechnical audits), but also (2) training and certification in standardized audit process and quality control procedures, similarly to the audit quality training required for financial auditors.

Research on algorithm auditing has so far called for auditor professionalization and certification as one of the requirements for high-quality audits [18, 54]. However, there is little discussion on *which* knowledge or training algorithm auditors should possess to perform audits effectively. Here, we advocate specifically for standard audit process and quality control procedures as a fundamental requirement for auditors performing criterion audits. In practice, this form of training and industry knowledge can leverage the vast existing body of work from the financial auditing industry, where standardized methods have been developed to support auditors in audit engagements—such as Auditing Standard (AS) 1105 on audit evidence [52], the International Standard on Auditing (ISA) 315 on risks of material misstatement [68], and the International Auditing and Assurance Standards Board's standard on quality management for auditing firms [69]). Furthermore, this training should also be complemented by comprehensive responsible AI education, such as on rigorous technical testing and effective risk management practices, to equip auditors with substantive domain knowledge when navigating complex considerations.

[16] The Delegated Regulation on independent auditing under the Digital Services Act [23] identifies three specific audit risks: (1) inherent risks, (2) control risks, and (3) detection risks.

4 ILLUSTRATION: ADAPTING THE CRITERION AUDIT FOR NYC BIAS AUDIT LAW

In November 2021, the New York City Council passed Local Law 144 of 2021 which requires bias audits for all automated employment decision tools ('AEDTs') used in hiring and promotion. The final rule, effective in July 2023, defines *bias audit* as "an impartial evaluation by an independent auditor," and the audit requires, as the minimum, an assessment of the tool's disparate impact on persons of any component 1 category—i.e., race/ethnicity and gender categories. Independence rules are also established, in which auditors are not allowed to have been involved in using, developing, and distributing the AEDT, or to have contractual and financial interest in the organizations using, developing, and distributing the AEDT.

4.1 Audit criteria development

Applying the criterion audit framework, we derived a set of criteria which aims to determine whether an algorithmic system has met the requirements of NYC bias audit law. The audit criteria were developed as a function of (1) the legislation's content and specifications, and (2) our practitioner experience in directly conducting technical bias testing and ethical risk and impact assessments on algorithmic systems. Moreover, the audit criteria are constructed using the previously discussed indirect verification approach (see Section 3.3).

The bias audit as stated in the legislation requires, at the minimum, a disparate impact analysis of the automated tool. Such technical analysis being the only explicit requirement for an audit stands in stark contrast to the broadly adopted view of current scholarship that mathematical-only perspectives of bias are insufficient in capturing impacts of algorithmic bias or in preventing discriminatory outcomes [58]. Moreover, the reviewed literature has documented extensively the limitations of a technical-only perspective in algorithmic bias and discrimination [34, 77], the importance of sociotechnical views of bias [4], the interdependence between technical and sociotechnical views in auditing of algorithmic systems [29], and the limitations of operationalizing technical measures in non-AI contexts [7, 19].

Adopting these sociotechnical views on bias management and mitigation, we designed our audit criteria to include three sections: (1) disparate impact analysis, (2) governance, and (3) risk assessment. For disparate impact analysis, we derived, to the best of our knowledge and ability, the minimally sufficient criteria allowing an auditor to evaluate a good faith analysis of the tool's disparate impact performed by the audited organization. For governance, a set of requirements aims at evaluating the existing governance body within the auditee's organization who is responsible and accountable for the management of risks related to bias of the AEDT. For risk assessment, we derived minimal requirements for an assessment of risks conducted by the auditee for the tool with a specific focus on bias.

4.2 Sections of an audit

Table 1 shows our set of audit criteria which aims to evaluate an algorithmic system for compliance with NYC bias audit law. The full set of criteria including all sub-criteria is available in Appendix A. These criteria provide the basis for both reviewing submitted documentation and verifying evidence (see 3.4). While the execution of this audit can be tackled by multiple auditors (e.g., based on their areas of domain expertise), the three audit sections are not intended to be independent but rather complementary, both conceptually and operationally, to each other. This holistic approach is designed to consider technical bias related risks in the sociotechnical context of the algorithm and governance measures.

4.2.1 Disparate impact analysis. The law requires the disclosure of the disparate impact assessment results in the form of impact ratios for groups of race/ethnicity, sex, and their intersections. An auditee has to have made a number of key

Table 1. Audit criteria for NYC Local Law 144 of 2021.

ID	Criterion
Q	**Disparate Impact Analysis**
Q.A	The tool analyzed for disparate impact shall be defined.
Q.B	The dataset based on which disparate impact is analyzed shall be defined and characterized.
Q.C	The demographic categories for which disparate impact can be analyzed using the dataset shall be defined.
Q.D	Where the selection rate method is used, positive and negative outcomes of the tool shall be clearly defined as the basis for selection rate.
Q.E	A metric which corresponds to selection rate or scoring rate shall be defined.
Q.F	The 'favored group' and 'disfavored groups' [47] shall be identified, for all demographic categories.
Q.G	The impact ratios shall be disclosed for all disfavored groups, for all demographic categories.
Q.H	Where the selection rate method was used, statistical significance calculations of the difference between selection rates shall satisfy Uniform Guidelines on Employee Selection Procedures (UGESP) [47].
G	**Governance**
G.A	The auditee shall have a party which is accountable for risks related to disparate impact.
G.B	The duties of the party accountable for risks related to disparate impact shall be clearly defined.
G.C	The auditee shall provide evidence that the defined duties of the party accountable for risks related to disparate impact are carried out.
R	**Risk Assessment**
R.A	The auditee shall have completed a risk assessment of the tool.
R.B	The risk assessment shall show identification of relevant risks related to bias.
R.C	The risk assessment shall demonstrate appropriate evaluation of relevant risks.

decision points to arrive at the quantitative results. These subjective and discretionary decision points mirror concepts known in the qualitative research as 'choice moments' [56]. Our criteria was accordingly designed to elucidate these decision points from the auditee and provide auditors a mechanism to evaluate their appropriateness.

The auditors evaluate: (1) the definition of AEDT by the auditee (Q.A), (2) the dataset used for analysis (Q.B), (3) the demographic information and its data collection process (Q.C), (4) the selection or scoring rate definition and its basis (Q.D & Q.E), (5) demographic groups covered in the analysis (Q.F), (6) impact ratio calculations, including of uncertainties (Q.G), and (7) statistical significance calculations of the difference between selection rates (Q.H).

For each criterion, a set of sub-criteria provides further guidance for auditors to assess the quality of the auditee decisions (see Appendix A). For instance, sub-criteria Q.D.1 requires the auditee to provide the justification for their choice of positive outcome as the basis for selection rate.[17] An auditor—having subject matter competency in a technical domain such as data science in combination with experience in algorithmic bias work—would be able to then evaluate whether, for example, using interview rates or hiring rates resulting from the AEDT use is more appropriate for calculations.

[17] 'Positive outcome' refers to the favorable outcome for a candidate from the use of the model, such as being selected to move forward in the hiring process or assigned a classification by an model.

4.2.2 Governance. Although the auditing of governance structures of an AEDT is not specified as requirements by the law, this set of criteria requires that the auditee has an accountable party for risks related to bias in a way that is clearly defined and operationalized across the organization.

Regulatory guidance and industry best practices have emphasized the role of governance and internal controls as a foundational building block of AI risk management systems [45, 48, 74]. Furthermore, our practitioner experience in performing audit-type engagements has also led us to believe that effective mitigation of the risks posed by an organization's use, development, and deployment of algorithm (including ones related to bias) requires some designated oversight body within the organization that is accountable for them. The rationale for including these criteria was therefore to determine whether this minimally sufficient layer of accountability is established.

The criteria require: (1) that the auditee has an accountable party for risks related to disparate impact (G.A), (2) that the responsibilities of this party are clearly defined (G.B), and (3) that such responsibilities have been carried out prior to the audit (G.C).

Similar to those in disparate impact analysis, specifications for the evidence in the form of sub-criteria require the auditee to show formalization of such governance. For instance, sub-criteria G.B.2 requires the auditee to show that the accountable party must have influence (e.g., through institutional power) over product changes. Evidence for this section can take the form of policies and procedures related to internal oversight—such as the charter of a responsible AI team or AI risk committee, their duties, and testimony by designated parties in the organization.

4.2.3 Risk assessment. Scholarship has highlighted that technical measures of risks of harm (such as bias) need to be understood in the sociotechnical context of an algorithmic system [1, 45, 61, 63]. This set of criteria examines the degree to which awareness for these sociotechnical risks is shown by the auditee.

Specifically, criteria in this section require (1) that the auditee has completed a risk assessment of their AEDT (R.A), (2) that risks have been identified (R.B), and evaluated (R.C), as part of this assessment. To satisfy these criteria, the auditee is required to show, in their risk assessment, sufficient awareness of sociotechnical risks of harm, and of bias in particular. Moreover, auditors tasked with evaluating these criteria can further verify the quality of this risk assessment in a verification interview where contributors of the risk assessment may be invited to speak on the details of the risks and their justifications.

4.3 Interactivity between sections of the audit

The three sections (i.e., disparate impact analysis, governance, and risk assessment) are intended to complement each other, reflecting the interwoven nature of the various aspects of AI risk management. Hasan et al. [29] provides an exemplary case for how this interplay between these areas should play out on the side of the audited entity when performing these auditable procedures. By identifying and evaluating risks related to bias, the auditee gains a sociotechnical understanding of their AEDT, which then informs whether and which risks can be measured using the available data, and how this technical testing should be performed. This dependency should, as a result, also be reflected in the disparate impact analysis, whose metrics and heuristics are guided by the risk assessment, and whose results reciprocally inform risk mitigation measures.[18]

We expect auditor evaluations of these components to also have this sense of interdependence. Consider the scenario where the auditee has identified, in their risk assessment, that a risk of bias is created by the way an automated system's

[18] Requirements of the EU AI Act for high-risk AI systems also reflect the importance of this interaction between technical testing and risk management, in which system providers are required to identify risk management measures through diligent system testing, disclose such testing heuristics, and justify their testing metrics.

user interface used in hiring is paginated such that candidates appearing on the first page is significantly more likely to receive an interview or be contacted (e.g., by recruiters). The auditor evaluating the risk assessment is encouraged, upon discovery of this information, to inform the auditor tasked with the disparate impact analysis of this information, so that they may take into account this detail when evaluating the basis of the metric used to calculate impact ratios. The disparate impact auditor should accordingly assess whether this element is taken into consideration in the technical analysis at all, and whether their metric is appropriate given the identified risks. Moreover, this procedural feature operationalizes the integration between what Mökander [42] refers to as 'narrow' and 'broad' auditing, where 'narrow' is exemplified by technology-focused testing or assessments and 'broad' is characterized by process-focused review of management of the technology.

5 BENEFITS & LIMITATIONS

5.1 Benefits

There are three key benefits to the proposed criterion audit:

5.1.1 Adaptability. Target scoping enables the audit to be tailored to a variety of algorithm use cases and legislative works. By stipulating minimally necessary and jointly sufficient conditions, the audit is a resource-efficient mechanism that can have a high degree of impact to provide assurance for a specific algorithmic system.

5.1.2 Efficiency and scalability. The proposed framework provides an efficient and scalable method for auditing. In many cases, organizations rely on dozens to hundreds of algorithms, which presents a particular challenge for effective governance and assessments. By performing AI audits in a systematic and transparent manner, our proposed approach ensures that a larger number of algorithms can be audited in a consistent and clearly defined set of procedures.

5.1.3 Transparency and accountability. Public disclosure of audit results and audit criteria provides a high degree of transparency. Considering the nascent stage of algorithm auditing as an industry, such level of transparency can be a powerful mechanism to foster accountability by (1) making the effectiveness of our proposed methodology measurable and comparable against other compliance and assurance frameworks, and (2) providing grounds for public scrutiny of AI companies concerning, among others, their adherence to regulations and ability to fulfill obligations to stakeholders and society.

5.2 Implementation challenges

There are a number of challenges in implementing the proposed audit framework:

5.2.1 Auditing standards development. Developing a set of standards precisely scoped to a legislative work is not a straightforward or unambiguous task. As stated in Section 3.3, audit criteria can be built using various approaches to auditing. In addition, in cases where substantive differences of opinion exist between experts—such as in the discourse on computational definitions of fairness [6, 8, 35]—designers of audit standards must navigate these nuances while also balancing to achieve a set of criteria that provides compliance and is operationalizable at scale.

Another issue focuses on *who* should be developing auditing standards. Standards are a powerful means to establish auditing norms and practices, and there are warranted concerns about corporate capture to shape them in accordance with their interests [79]. In financial auditing, the International Financial Reporting Standards (IFRS) are developed and maintained by the independent International Auditing and Assurance Standards Board (IAASB). By contrast, auditing criteria for AI regulation compliance are currently emerging, where sets of criteria, despite being developed by independent standard-setting bodies [12, 14, 31, 33], have yet gained wide adoption from the industry.

5.2.2 Training and gray area decision-making. Following the training and accreditation of financial auditors, emphasis must be placed on equipping auditors with the competency and capabilities to perform these audits. As critical judgment is a fundamental feature of this methodology, auditors need to be equipped to better understand whether and, if so, the extent to which a given criteria may or may not be sufficiently met, or if further consultation is required. However, there is currently little discourse in the algorithm auditing field and limited training resources for aspiring algorithm auditors on this matter. The current state of auditor training unfortunately relies largely on on-the-job training, and self-empowerment for making judgment calls.

5.2.3 Opinion shopping and its effect on auditor independence. As the AI auditing industry is in its infancy, there lacks a set of safeguards sufficiently robust to ensure high quality audits. Currently, there is little to none preventing auditors from making their audits as easy to 'pass' as possible to capture the market of organizations in scope for audits. This lack of safeguards enables audited entities to search for auditors who are more likely to provide a favorable opinion or whose audits are less stringent, known as *opinion shopping* in the financial sector [15, 37, 39]. This practice creates an environment in which self-imposed substantial independence requirements for algorithm auditors become a business risk for auditing firms. In the case of NYC bias audit law, auditors are thus incentivized to maintain only the minimum independence requirements specified by the regulation, *irrespective* of whether such specifications are sufficiently robust against deteriorating audit quality. This superficial sense of rigor can greatly perpetuate 'audit-washing,' whereby harms supposedly prevented by the audit are instead distracted from or even excused [27]. This problem is in contrast to that in the financial sector, in which strong independence rule has two hundred years of history, precedent, and significant support from various stakeholders in the auditing ecosystem, including regulatory authorities.

5.3 Limitations

The criterion audit framework has a number of intrinsic limitations:

5.3.1 Reliance on the regulatory requirements. The criterion audit is by design dependent on the formal requirements of a regulation or other standardized frameworks. Effective rulemaking and precise scoping of the requirements allow room for the audit and its criteria to become instrumental in meaningfully preventing or mitigating algorithmic harms. On the one hand, excessively broad scope risks the criterion audit not having enough teeth to enforce meaningful changes to the status quo. On the other hand, if the legal requirements are excessively restrictive, the criteria risk becoming a checklisting exercise for auditors, failing to achieve the objectives set forth by the law.[19] For NYC bias audit law, the rule specifies various details on disclosure of impact ratios as a metric to assess disparate impact but does not require actions to address, manage, and mitigate bias and discrimination harms resulting from such tools in a systematic manner. This misdirected focus restricts the scope of the criteria greatly and thereby limits its ability to fulfill its intended goals. Here, we urge regulators to pay close attention to the practitioner experience of auditors to

[19] Policymakers face a similar challenge for the level of precision in rulemaking: if regulations are too specific, they cannot be appropriate for or adaptable to every given system; on the other hand, if too broad or general, they risk too easily passing the requirements.

understand to what extent the goals of their proposed rules can be met within the bounds of auditing mechanisms such as the criterion audit.

5.3.2 Reliance on an audit oversight ecosystem. Modeled after financial auditing, this audit framework similarly requires a network of entities working in tandem to ensure its effectiveness. This network includes various stakeholders, from auditing firms, audited entities, entities developing standards, to entities providing training and certifying practitioners. Overseeing this network requires not only lawmakers, but also regulatory bodies who have the authority to appoint independent parties, certify standards, and enforce the law. As this ecosystem is currently not fully developed, audits performed using this audit framework may yet suffer from audit quality issues.

5.3.3 Reliance on a clearly defined outcome taxonomy. In financial auditing, an audit outcome is expressed as an opinion by the auditor indicating whether the financial statements are presented fairly. There are four types of audit opinion; each one has a clear definition about the nature of the financial statements, and specifies what users should expect from the auditor in the audit report. A 'disclaimer of opinion,' for example, requires an auditor to explain why an opinion is withheld and to explicitly indicate that no opinion is expressed. Based on this information, the users of the audit report (e.g., shareholders) are thereby enabled to make appropriate financial decisions.

This audit framework requires a similar set of outcome taxonomies, but there is no established or widely adopted equivalent in the AI auditing industry, at least to the extent that it meaningfully enables readers of the audit report to take appropriate actions. More concerning is that some regulations do not specify the disclosure of any audit outcome at all.[20] Such is the case for NYC bias audit law, which does not require the automated tools to even *pass* the audit to achieve compliance.[21] Moreover, without established industry norms around how audits ought to be used to make informed decisions, even attempts to implement a simple pass/fail opinion on audits may run the risk of diluting the significance of such judgment and rendering it difficult to interpret for audit stakeholders.

6 CONCLUSION

In this paper, we have proposed the criterion audit framework for external compliance and assurance audits of AI systems. In the absence of standard practices for audits as proposed by emerging regulations, our approach offers an operationalizable auditing methodology aiming to provide assurance to audit stakeholders about organizations' ability to fulfill obligations about their algorithms. We lay out the necessary conditions for an effective criterion audit: an evaluation of an algorithmic system (1) against transparent audit criteria, (2) performed by qualified professionals, (3) with partially disclosed results to the public (4) about whether it complies with the requirements of a legislative work or standard. We offer procedures to perform the engagement for auditors, and highlight their responsibilities as evaluators of evidence about AI systems and the need for specialized training. We adapt the framework to NYC bias audit law, whereby we show the audit criteria to perform the audit, and provide the rationale and illustrative examples on auditor evaluation for each audit section, and the interplay between auditors tasked with different audit sections. We conclude by examining the benefits, implementation challenges, and inherent limitations of this audit framework, in part drawing from our perspective as auditors conducting bias audits under the NYC bias audit law. Our paper provides a glimpse into how auditing for compliance and assurance can fall short in a nascent industry even when modeled after practices of a more mature one. Our discussion emphasizes the critical need for an ecosystem—comprising auditors,

[20] The only exception is the Digital Services Act, in which the Delegated Regulation on independent auditing [23] specifies the audit outcomes as either 'positive,' 'positive with comments,' or 'negative,' corresponding to the auditee's extent of compliance.

[21] Compliance can simply be met when a tool, prior to its use: (1) has a bias audit conducted within one year, and (2) has a result summary of the most recent audit be made publicly available.

auditees, regulators, standard-setting bodies, certification bodies, and enforcing agencies—working in tandem to ensure high-quality compliance and assurance external audits. Finally, even when such an ecosystem matures, our proposal for auditing (and more generally, AI auditing as a practice) is only one piece in the responsible AI puzzle, and—much similar to financial auditing—is not the answer to all problems of accountability.

7 ETHICS STATEMENT

7.0.1 Positionality statement. All authors of this paper work for a for-profit business that has been using this audit framework to perform bias audits under NYC Local Law 144 of 2021 since 2022. This work is thus drawn on our perspectives (e.g., values, norms, practices, and biases, explicitly and implicitly) (1) as consultants providing advisory services, (2) as auditors conducting algorithm audits for companies using and developing AI, and (3) as researchers of responsible AI and AI governance. More specifically, our discussion and prioritization on benefits, limitations, and implementation challenges draw heavily on our experiences having used this audit framework to perform the bias audits. They are, however, not intended to be comprehensive of all audit stakeholders' perspectives (e.g., the public, audit report readers, regulators, or auditees).

7.0.2 Adverse impact statement. We recognize that there are risks associated with the misuse of our methodology. For example, our criteria for NYC bias audit law was designed assuming certain ways that the AI system is built. While we initially tested the criteria's applicability using publicly available information about AI hiring systems, we could not have imagined all the possible ways hiring algorithms are designed, developed, deployed, and tested. Uncritically using our criteria to perform audits can thus run into issues concerning adaptability to real-world hiring systems. In addition, our criteria can also be 'reverse engineered' by companies to game our audit process, although we expect this risk to be of no immediate cause for concern. More generally, there are also risks of misinterpretation of our analysis. Our operationalizable audit framework can be seen as conveying a false sense of what auditing can do to advance accountability without situating it the appropriate context—i.e., that it is an instrument whose effectiveness requires the existence of a supporting multi-stakeholder ecosystem. We also discuss such issues in section 5.3 and 5.2.

7.0.3 Ethical considerations statement. Due to the nature of our research, we did not undergo any Institutional Review Board (IRB) process. However, as we acknowledge our position as one where audits being widely adopted is a business interest, our discussion is intended to only critically reflect on the ways auditing may benefit the current discourse on AI transparency and accountability.

ACKNOWLEDGMENTS

We would like to thank the Notre Dame-IBM Tech Ethics Lab for supporting this work. Notre Dame-IBM Tech Ethics Lab had no role in the design and conduct of the study; the access and collection of data; the analysis and interpretation of data; the preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. All authors work for BABL AI.

REFERENCES

[1] Ada Lovelace Institute and DataKind UK. 2020. *Examining the Black Box: Tools for Assessing Algorithmic Systems.* Technical Report. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

[2] Ifeoma Ajunwa. 2021. An Auditing Imperative for Automated Hiring. *Harvard Journal of Law & Technology* 34, 2 (06 2021), 80 pages. https://doi.org/10.2139/ssrn.3437631

[3] Julia Angwin and Surya Mattu. 2016. Amazon Says It Puts Customers First. But Its Pricing Algorithm Doesn't. *ProPublica* (09 2016). https://www.propublica.org/article/amazon-says-it-puts-customers-first-but-its-pricing-algorithm-doesnt

[4] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (08 2016), 671–732. https://doi.org/10.2139/ssrn.2477899

[5] Kathy Baxter. 2021. *AI Ethics Maturity Model.* Technical Report. Salesforce AI Research. https://www.salesforceairesearch.com/static/ethics/EthicalAIMaturityModel.pdf

[6] Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. 2023. The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 400–422. https://doi.org/10.1145/3593013.3594007

[7] Mario Biagioli. 2016. Watch out for Cheats in Citation Game. *Nature* 535 (07 2016), 201. https://doi.org/10.1038/535201a

[8] Reuben Binns. 2020. On the Apparent Conflict between Individual and Group Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 514–524. https://doi.org/10.1145/3351095.3372864

[9] Board of Governors of the Federal Reserve System. 2011. SR 11-7: Guidance on Model Risk Management. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

[10] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The Algorithm Audit: Scoring the Algorithms That Score Us. *Big Data & Society* 8, 1 (01 2021). https://doi.org/10.1177/2053951720983865

[11] Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Graham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. (04 2020). arXiv:2004.07213 [cs.CY]

[12] Bundesamt für Sicherheit in der Informationstechnik (BSI). 2022. AI Cloud Service Compliance Criteria Catalogue (AIC4). https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/AIC4/aic4_node.html

[13] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). Proceedings of Machine Learning Research, New York, NY, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[14] Ryan Carrier, Shea Brown, Merve Hickok, Cari Miller, Michael McCarthy, Esther Chung, Joshua Scarpino, Heidi Saas, and Marc Hébert. 2021. *New York City Local Law 144: Bias Audits for Automated Employment Decision Tools.* Technical Report. ForHumanity. https://forhumanity.center/web/wp-content/uploads/2023/10/New-York-City-Bias-Audit-An-Overview-and-Action-Plan-v2.pdf

[15] Jong-Hag Choi, Heesun Chung, Catherine Heyjung Sonu, and Yoonseok Zang. 2018. Opinion Shopping to Avoid a Going Concern Audit Opinion and Subsequent Audit Quality. *Auditing: A Journal of Practice & Theory* 38, 2 (05 2018), 101–123. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3182103

[16] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). Proceedings of Machine Learning Research, New York, NY, USA, 134–148. https://proceedings.mlr.press/v81/chouldechova18a.html

[17] Yvette D. Clarke. 2022. H.R. 6580 – Algorithmic Accountability Act of 2022. https://www.congress.gov/bill/117th-congress/house-bill/6580

[18] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1571–1583. https://doi.org/10.1145/3531146.3533213

[19] Jón Daníelsson. 2002. The Emperor Has No clothes: Limits to Risk Modelling. *Journal of Banking & Finance* 26, 7 (07 2002), 1273–1296. https://doi.org/10.1016/s0378-4266(02)00263-7

[20] Alex Engler. 2021. Auditing Employment Algorithms for Discrimination. *The Brookings Institution* (03 2021). https://www.brookings.edu/articles/auditing-employment-algorithms-for-discrimination/

[21] European Commission. 2021. Proposal for a Regulation of the European Parliament and the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. https://assets-global.website-files.com/637e4725db842e4068de0899/6565f4809623754436366a2b_COMMISSION%20PROPOSAL.pdf

[22] European Parliament and Council of the European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065

[23] European Parliament and Council of the European Union. 2023. Delegated Regulation on Independent Audits under the Digital Services Act. https://digital-strategy.ec.europa.eu/en/library/delegated-regulation-independent-audits-under-digital-services-act

[24] Federal Trade Commission. 2022. Trade Regulation Rule on Commercial Surveillance and Data Security. https://www.federalregister.gov/documents/2022/08/22/2022-17752/trade-regulation-rule-on-commercial-surveillance-and-data-security

[25] ForHumanity. 2022. ForHumanity University. https://forhumanity.center/forhumanity-university/

[26] ForHumanity. 2023. *Audit Manual for Independent Audit of AI Systems v1.5*. https://forhumanity.center/web/wp-content/uploads/2023/08/ForHumanity-IAAIS-Audit-Manual

[27] Ellen P. Goodman and Julia Trehu. 2023. Algorithmic Auditing: Chasing AI Accountability. *Santa Clara High Technology Law Journal* 39, 3 (05 2023), 289–338. https://digitalcommons.law.scu.edu/chtlj/vol39/iss3/1

[28] James Guszcza, Iyad Rahwan, Will Bible, Manuel Cebrian, and Vic Katyal. 2018. Why We Need to Audit Algorithms. *Harvard Business Review* (11 2018). https://hbr.org/2018/11/why-we-need-to-audit-algorithms

[29] Ali Hasan, Shea Brown, Jovana Davidovic, Benjamin Lange, and Mitt Regan. 2022. Algorithmic Bias and Risk Assessments: Lessons from Practice. *Digital Society* 1, 2 (08 2022), 20 pages. https://doi.org/10.1007/s44206-022-00017-z

[30] Marc P. Hauer, Lena Müller-Kress, Gertraud Leimüller, and Katharina Zweig. 2023. Using Assurance Cases to Assure the Fulfillment of Non-functional Requirements of AI-based Systems – Lessons Learned. In *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. Institute of Electrical and Electronics Engineers (IEEE), Dublin, Ireland, 172–179. https://doi.org/10.1109/ICSTW58534.2023.00040

[31] Institute of Electrical and Electronics Engineers (IEEE). 2021. IEEE Ontological Standard for Ethically Driven Robotics and Automation Systems. *IEEE Std 7007-2021* (2021), 1–119. https://doi.org/10.1109/IEEESTD.2021.9611206

[32] Institute of Electrical and Electronics Engineers (IEEE). 2023. IEEE CertifAIEd Authorized Assessor Training. https://www.ethosai.ai/home-old-bj5

[33] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC). 2023. *ISO/IEC 42001:2023 – Information Technology – Artificial Intelligence – Management System* (1st ed.). International Standard. https://www.iso.org/standard/81230.html

[34] Pauline Kim. 2017. Auditing Algorithms for Discrimination. *University of Pennsylvania Law Review Online* 166, 17-12-03 (12 2017), 189–203. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3093982

[35] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[36] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (05 2016). https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[37] Clive Lennox. 2000. Do Companies Successfully Engage in Opinion-shopping? Evidence from the UK. *Journal of Accounting and Economics* 29, 3 (06 2000), 321–337. https://doi.org/10.1016/s0165-4101(00)00025-2

[38] Laura Lucaj, Patrick van der Smagt, and Djalel Benbouzid. 2023. AI Regulation Is (Not) All You Need. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1267–1279. https://doi.org/10.1145/3593013.3594079

[39] David McCann. 2019. 'Opinion-Shopping' Compromises Auditor Independence. *CFO.com* (05 2019). https://www.cfo.com/news/opinion-shopping-compromises-auditor-independence/657865/

[40] Phil Mendelson. 2021. B24-0558 – Stop Discrimination by Algorithms Act of 2021. https://legiscan.com/DC/bill/B24-0558/2021

[41] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-Construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 735–746. https://doi.org/10.1145/3442188.3445935

[42] Jakob Mökander. 2023. Auditing of AI: Legal, Ethical and Technical Approaches. *Digital Society* 2, 49 (2023), 32 pages. https://doi.org/10.1007/s44206-023-00074-y

[43] Jakob Mökander, Maria Axente, Federico Casolari, and Luciano Floridi. 2022. Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines* 32 (2022), 241–268. https://doi.org/10.1007/s11023-021-09577-4

[44] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2023. Auditing Large Language Models: A Three-Layered Approach. *SSRN Electronic Journal* (02 2023), 29 pages. https://doi.org/10.2139/ssrn.4361607

[45] National Institute of Standards and Technology (NIST). 2023. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. https://doi.org/10.6028/nist.ai.100-1

[46] Office of Federal Contract Compliance Programs (OFCCP). 1978. 41 CFR 60-3.4: Information on Impact. https://www.ecfr.gov/current/title-41/part-60-3/section-60-3.4

[47] Office of Federal Contract Compliance Programs (OFCCP). 2019. Validation of Employee Selection Procedures: How does OFCCP Identify Disparities (Adverse Impact) Caused by Use of Employee Selection Procedures? https://www.dol.gov/agencies/ofccp/faqs/employee-selection-procedures#Q4

[48] Office of the Comptroller of the Currency (OCC). 2021. *Model Risk Management*. Comptroller's Handbook. https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html

[49] Frank Pallone. 2022. H.R.8152 – American Data Privacy and Protection Act. https://www.congress.gov/bill/117th-congress/house-bill/8152

[50] Michael Power. 1994. *The Audit Explosion*. Technical Report. Demos, London, UK. https://www.demos.co.uk/files/theauditexplosion.pdf

[51] Michael Power. 1999. *The Audit Society: Rituals of Verification*. Oxford University Press, Oxford, UK. https://doi.org/10.1093/acprof:oso/9780198296034.001.0001

[52] Public Company Accounting Oversight Board (PCAOB). 2022. AS 1105: Audit Evidence. https://pcaobus.org/oversight/standards/auditing-standards/details/AS1105

[53] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-end Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

[54] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 557–571. https://doi.org/10.1145/3514094.3534181

[55] Paul Sarbanes and Michael Oxley. 2002. Sarbanes-Oxley Act of 2002. https://sarbanes-oxley-act.com

[56] Maggi Savin-Baden and Claire Howell Major. 2012. *Qualitative Research: The Essential Guide to Theory and Practice* (1st ed.). Routledge, Oxfordshire, UK.

[57] Jonas Schuett. 2022. Three Lines of Defense against Risks from AI. , 22 pages. arXiv:2212.08364 [cs.CY]

[58] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* NIST Special Publication 1270. National Institute of Standards and Technology (NIST). https://doi.org/10.6028/nist.sp.1270

[59] Securities and Exchange Commission (SEC). 1972. 17 CFR 210.2-01: Qualifications of Accountants. https://www.ecfr.gov/current/title-17/part-210/section-210.2-01

[60] Andrew D. Selbst. 2021. An Institutional View of Algorithmic Impact Assessments. *Harvard Journal of Law & Technology* 35, 1 (06 2021), 117–191. https://papers.ssrn.com/abstract=3867634

[61] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[62] Mona Sloane. 2021. The Algorithmic Auditing Trap. *OneZero* (03 2021). https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d

[63] Mona Sloane and Emanuel Moss. 2023. Assessing the Assessment: Comparing Algorithmic Impact Assessments and AI Audits. *SSRN Electronic Journal* (06 2023), 14 pages. https://doi.org/10.2139/ssrn.4486259 In review for edited volume for Oxford University Press.

[64] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots. *American Civil Liberties Union* (07 2018). https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28

[65] The Algorithmic Bias Lab. 2022. AI and Algorithm Auditor Certificate Program. https://courses.babl.ai/p/ai-and-algorithm-auditor-certification

[66] The International Auditing and Assurance Standards Board (IAASB). 2013. *International Standard on Assurance Engagements (ISAE) 3000 Revised, Assurance Engagements Other Than Audits or Reviews of Historical Financial Information.* Standards and Pronouncements 978-1-60815-167-7. The International Federation of Accountants (IFAC). https://www.iaasb.org/publications/international-standard-assurance-engagements-isae-3000-revised-assurance-engagements-other-audits-or

[67] The International Auditing and Assurance Standards Board (IAASB). 2014. *A Framework for Audit Quality: Key Elements That Create an Environment for Audit Quality.* Standards and Pronouncements 978-1-60815-178-3. The International Federation of Accountants (IFAC). https://www.iaasb.org/publications/framework-audit-quality-key-elements-create-environment-audit-quality-3

[68] The International Auditing and Assurance Standards Board (IAASB). 2019. *ISA 315 (Revised 2019): Identifying and Assessing the Risks of Material Misstatement.* Standards and Pronouncements. The International Federation of Accountants (IFAC). https://www.iaasb.org/publications/isa-315-revised-2019-identifying-and-assessing-risks-material-misstatement

[69] The International Auditing and Assurance Standards Board (IAASB). 2020. *International Standard on Quality Management (ISQM) 1: Quality Management for Firms That Perform Audits or Reviews of Financial Statements, or Other Assurance or Related Services Engagements.* Handbooks, Standards, and Pronouncements. The International Federation of Accountants (IFAC). https://www.iaasb.org/publications/international-standard-quality-management-isqm-1-quality-management-firms-perform-audits-or-reviews

[70] The International Auditing and Assurance Standards Board (IAASB). 2022. *2022 Handbook of the International Code Of Ethics for Professional Accountants.* Handbooks 978-1-60815-508-8. The International Federation of Accountants (IFAC). https://www.ethicsboard.org/publications/2022-handbook-international-code-ethics-professional-accountants

[71] The New York City Council. 2021. Subchapter 25: Automated Employment Decision Tools. https://codelibrary.amlegal.com/codes/newyorkcity/latest/NYCadmin/0-0-0-135598

[72] The New York City Council. 2023. Subchapter T: Automated Employment Decision Tools. https://codelibrary.amlegal.com/codes/newyorkcity/latest/NYCrules/0-0-0-138391

[73] U.S. Department of Health & Human Services. 2021. *Trustworthy AI (TAI) Playbook.* Technical Report. https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

[74] U.S. Government Accountability Office (GAO). 2021. *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities.* Technical Report GAO-21-519SP. https://www.gao.gov/products/gao-21-519sp

[75] Ben Wagner and Lubos Kuklis. 2021. *Establishing Auditing Intermediaries to Verify Platform Data.* Oxford University Press, Oxford, UK, Chapter 9, 169–179. https://doi.org/10.1093/oso/9780197616093.003.0010

[76] Sheridan Wall and Hilke Schellmann. 2021. We Tested AI Interview Tools. Here's What We Found. *MIT Technology Review* (07 2021). https://www.technologyreview.com/2021/07/07/1027916/we-tested-ai-interview-tools/

[77] Elizabeth Anne Watkins, Michael McKenna, and Jiahao Chen. 2022. *The Four-Fifths Rule Is Not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness.* Technical Report P22-1-v0.2.2. Parity Technologies, Inc. https://doi.org/10.48550/arXiv.2202.09519

[78] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 666–677. https://doi.org/10.1145/3442188.3445928

[79] Meg Young, Michael Katell, and P. M. Krafft. 2022. Confronting Power and Corporate Capture at the FAccT Conference. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1375–1386. https://doi.org/10.1145/3531146.3533194

A FULL AUDIT CRITERIA SET

Table 2. Full set of audit criteria for NYC Local Law 144 of 2021.

ID	*Criterion* & Sub-criterion
Q	**Disparate Impact Analysis**
Q.A	*The tool analyzed for disparate impact shall be defined.*
Q.A.1	Where the tool comprises more than one automated component, evidence shall show appropriate definition of the tool.
Q.B	*The dataset based on which disparate impact is analyzed shall be defined and characterized.*
Q.B.1	Evidence shall show justification for why this dataset is appropriate for analysis.
Q.B.2	Where test data as defined in §5-300 [72] was used, evidence shall show: (a) justification for not using historical data, (b) that the sample size of historical data is not sufficiently large to perform a statistically significant disparate impact analysis, and (c) the methodology by which the test data was collected.
Q.B.3	Evidence shall show: (a) that the most recent analysis was conducted less than one year prior to the start date of this audit, or after a major update to the tool, unless such update was more than one year prior to the start date of this audit, in which case, evidence shall show: (b) justification for why such analysis is still appropriate for this audit.
Q.B.4	Evidence shall show that the time span of the dataset is within one year of the start date of the analysis.
Q.C	*The demographic categories for which disparate impact can be analyzed using the dataset shall be defined.*
Q.C.1	Evidence shall identify the demographic categories for which disparate impact can be analyzed.
Q.C.2	Evidence shall show that such demographic categories include at the minimum: race/ethnicity and gender.
Q.C.3	Evidence shall disclose the method by which demographic data was collected.
Q.C.4	Evidence shall identify and disclose the demographic categories that are out of scope for this analysis.
Q.C.5	Where demographic data were inferred, evidence shall: (a) identify the method by which demographic data was inferred, and (b) show justification for why the selected method of demographic inference was appropriate.

Q.D *Where the selection rate method is used, positive and negative outcomes of the tool shall be clearly defined as the basis for selection rate.*

Q.D.1 Evidence shall show justification for why this definition of positive outcome is appropriate.

Q.D.2 Where thresholding is used as a basis for positive outcome determination, evidence shall show justification for why the level(s) of threshold is (are) appropriate.

Q.D.3 Evidence shall identify and disclose:

(a) all user-configurable tool settings,
(b) whether each setting affects positive outcomes,

and for all settings identified as outcome-affecting:

(c) their extents of user configurability,
(d) their default values, and
(e) justification for why such default values were appropriate.

Q.D.4 Evidence shall disclose the user-configurable tool settings and combinations of settings used for the analysis.

Q.E *A metric which corresponds to selection rate or scoring rate shall be defined.*

Q.E.1 Where the selection rate method is used, evidence shall show that the selection rate of a group is defined as the ratio of positive outcome to all outcomes for that group.

Q.E.2 Where the scoring rate method is used, evidence shall show that the scoring rate of a group is defined as the rate at which that group receives a score from the tool above the median score of the sample.

Q.F *The 'favored group' and 'disfavored groups' shall be identified, for all demographic categories.*[22]

Q.F.1 Evidence shall show that the favored and disfavored groups are identified based on selection rates or scoring rates.

Q.F.2 Evidence shall show that the groups pertaining to race/ethnicity satisfy §60-3.4 B of the EEO guidelines [46].

Q.F.3 Where the groups pertaining to race/ethnicity do not satisfy EEO guidelines, evidence shall show:

(a) justification for why such EEO grouping is not used, and
(b) the appropriateness of any substituted grouping.

Q.F.4 Evidence shall show that the groups pertaining to gender contains at least Male and Female.

Q.F.5 Evidence shall show that intersectional groups contain all permutations of race/ethnicity and gender groups.

Q.F.6 Where race/ethnicity and gender groups are not known for a sample of candidates assessed by the tool, the evidence shall disclose its sample size.

Q.G *The impact ratios shall be disclosed for all disfavored groups, for all demographic categories.*

Q.G.1 Where an impact ratio for a disfavored group is below 0.8, evidence shall show justification for why the disfavored group is disadvantaged.

Q.G.2 Evidence shall show results of uncertainty analysis (e.g., standard error for the mean) or error propagation of impact ratios in the form of absolute errors or error bars.

Q.G.3 Where demographic data was inferred, evidence shall show that systematic errors due to demographic inference are properly propagated in impact ratio calculations.

[22] 'Favored group' refers to the group with the highest selection or scoring rate in the analysis, and 'disfavored groups' refer to all other groups in the demographic category (i.e., with lower selection or scoring rate) [47].

Q.G.4 Where a gender, race/ethnicity, or intersectional group was excluded from impact ratio calculations due to its size being below 2% of the total sample size, evidence shall show:

(a) justification for its exclusion,
(b) its sample size, and
(c) its selection rate or scoring rate.

Q.H Where the selection rate method was used, statistical significance calculations of the difference between selection rates shall satisfy Uniform Guidelines on Employee Selection Procedures (UGESP) [47].

Q.H.1 Evidence shall show that statistical significance is calculated using the Two Independent-Sample Binomial Z-Test for sample sizes of 30 or more, and using the Fisher's Exact Test for sample sizes of fewer than 30.

G Governance

G.A The auditee shall have a party which is accountable for risks related to disparate impact.

G.A.1 Evidence should show that the accountable party is a committee, but may also show that the accountable party is a single individual.

G.A.2 Evidence shall clearly show that risks related to disparate impact are owned and managed by the accountable party.

G.B The duties of the party accountable for risks related to disparate impact shall be clearly defined.

G.B.1 Evidence shall show that such duties pertain to the ownership, management, and monitoring of risks related to disparate impact.

G.B.2 Evidence shall show that the accountable party has influence over product changes per effective challenge in Federal Guidance on Model Risk Management [9].

G.C The auditee shall provide evidence that the defined duties of the party accountable for risks related to disparate impact are carried out.

G.C.1 Evidence shall show that the defined duties were carried out prior to the start date of this audit.

R Risk Assessment

R.A The auditee shall have completed a risk assessment of the tool.

R.A.1 Evidence shall show that a risk assessment or an equivalent analysis was completed less than one year prior to the issuance date of this audit.

R.B The risk assessment shall show identification of relevant risks related to bias.

R.B.1 Evidence shall show the identification of risks related to various biases along all stages of the AI lifecycle, such as listed in the National Institute of Standards and Technology (NIST) Standard for Identifying and Managing Bias in Artificial Intelligence [58].

R.B.2 Evidence shall show awareness of the parties potentially affected by the decisions made along all stages of the AI lifecycle.

R.C The risk assessment shall demonstrate appropriate evaluation of relevant risks.

R.C.1 Evidence shall show that the identified risks are assessed from the perspectives of multiple affected external and internal stakeholders, with justifications for the extent of and mechanism by which such risks affect these stakeholders.

R.C.2 Evidence shall show that the identified risks are assessed:

(a) in a sufficiently rigorous manner, using a quantitative and/or qualitative evaluation scheme, and

(b) along multiple dimensions, such as but not limited to likelihood of harm and severity of harm.

R.C.3 Evidence shall show justification for the provided evaluation of risks.