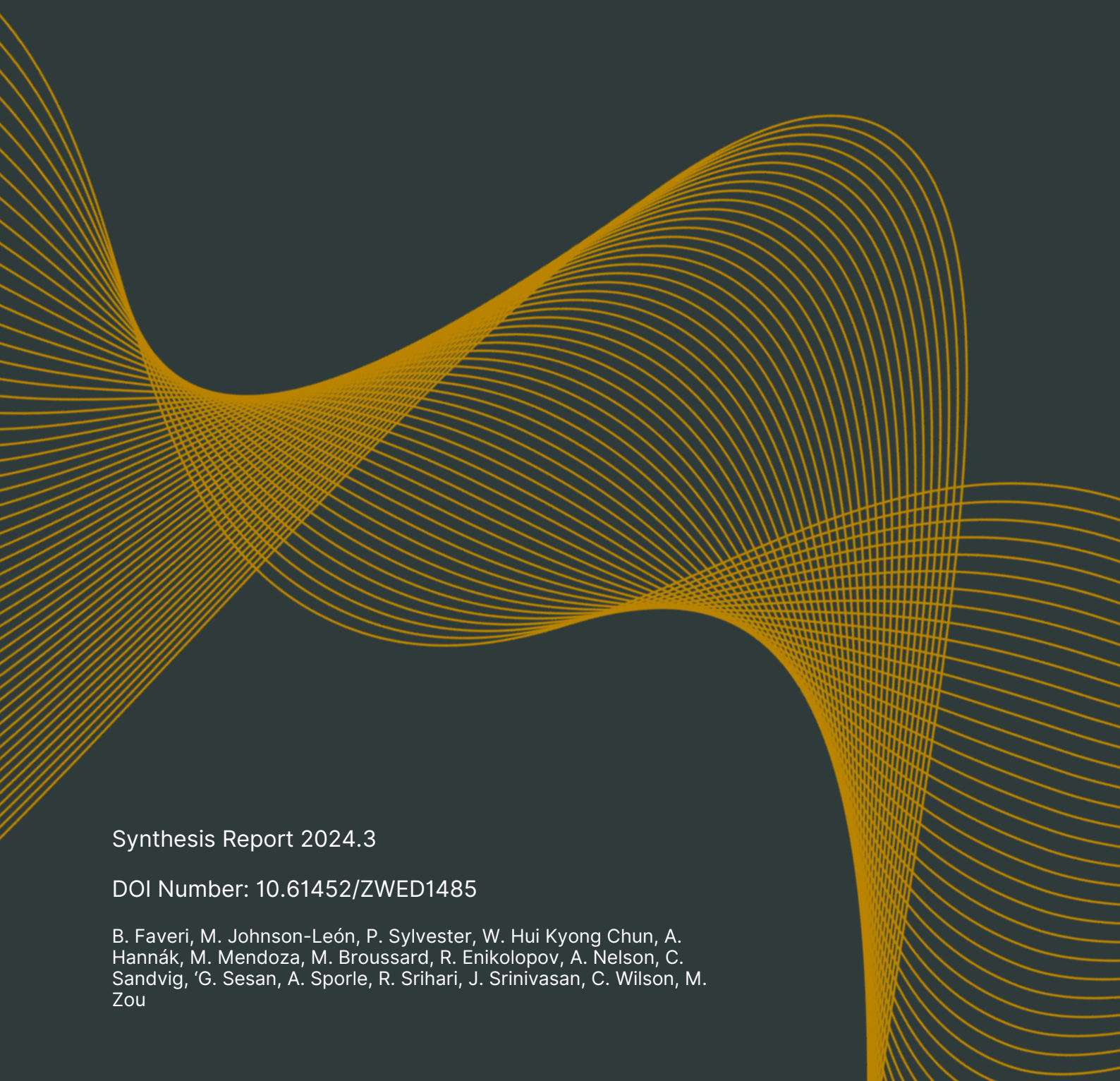
IPIE

International Panel on the Information Environment

Towards A Global AI Auditing Framework

Assessment and Recommendations



Synthesis Report 2024.3

DOI Number: 10.61452/ZWED1485

B. Faveri, M. Johnson-León, P. Sylvester, W. Hui Kyong Chun, A. Hannák, M. Mendoza, M. Broussard, R. Enikolopov, A. Nelson, C. Sandvig, ‘G. Sesan, A. Sporle, R. Srihari, J. Srinivasan, C. Wilson, M. Zou

IPIE
International Panel on the Information Environment

Towards A Global AI Auditing Framework

Assessment and Recommendations

Synthesis Report 2024.3

SYNOPSIS

The growing integration of artificial intelligence (AI) into critical sectors of society, from healthcare to education, has the potential to support widespread social transformation and progress. However, AI systems also have the power to perpetuate biases, deepen inequalities, and cause environmental harm.

Accurately evaluating the risks and benefits of an AI system requires a careful audit. Current approaches to auditing, however, rarely involve independent auditors, provide sufficient evidence, or account for global impacts.

Policymakers urgently need a comprehensive global framework for AI audits that validates genuine benefits and risks. The IPIE's *Scientific Panel on Global Standards for AI Audits* set out to independently establish global, cross-disciplinary scientific consensus on what makes an audit effective and trustworthy. The Panel consisted of 16 experts from computer science, the social sciences, and the humanities, with expertise spanning generative AI systems, algorithmic auditing, indigenous data sovereignty, data journalism, and the relationship between civil society, human rights, and AI.

This Synthesis Report is the Panel's second assessment and offers a technical explanation of why a global framework for AI audits must provide:

- common standards and procedures,
- auditor independence and qualifications,
- auditor access to evidence,
- transparent audit results, and
- evaluation of AI's disparate global impacts.

The proposed framework for AI audits addresses seven features:

1. audit scope,
2. auditor role,
3. mechanisms for stakeholder engagement,
4. context-appropriate audit criteria,
5. evidence requirements,
6. audit methodologies, and
7. comprehensive post-audit activities.

A global framework must take care to map relevant stakeholders, develop participatory methods, establish criteria to minimize potential harms, develop audit methodologies, and check if the audit meets stakeholder expectations.

CONTENTS

1. INTRODUCTION AND BACKGROUND ON AI AUDITS

The growing integration of artificial intelligence (AI) into critical sectors of society, from healthcare to education, has the potential to support widespread social transformation and progress. However, AI systems also have the power to perpetuate biases, deepen inequalities, and cause environmental harm. For example, AI models used in clinical settings in Thailand have at times produced inaccurate outputs due to local infrastructural conditions [1]. Vast quantities of training data, essential for AI systems, are annotated by underpaid workers in Kenya [2]. Platform companies using AI-calculated delivery times force gig workers into unhealthy work schedules [3]. Manufacturing graphical processing units, or GPUs, which are essential for the latest AI systems, results in massive emissions of greenhouse gases, accelerating global climate change [4]. These effects occur across many geographies where AI systems are developed, deployed, and used, and they are found across AI's supply chain and life cycle [5].

Public agencies have responded through guidelines and regulations, such as the EU's AI Act and international standards such as the ISO 42001:2023 [6],[7],[8]. Civil society groups have also pushed for AI systems to be developed and operated in fair and just ways that provide for disclosure and certification mechanisms that demonstrate how AI systems are developed and deployed in the public interest [9],[10]. Moreover, some developers and deployers of AI systems have been eager to find ways to assess whether their systems adhere to certification standards and ethical principles. For all these actors, audits have quickly become an important tool for evaluating AI systems.

Auditing an AI system can help evaluate its interactions with individuals, communities, and organizations and assess whether these systems are properly developed, deployed, operated, and managed. An audit can check whether an AI system adheres to vital social, ethical, and legal norms such as fairness, data privacy, and environmental sustainability. For example, audits can be used to evaluate whether racialized and poorer tenants have been forced to pay

higher premiums on housing insurance because of the use of algorithmic systems [11]. When an audit detects deviations from those conditions it can help demonstrate the risks and harms that accrue from AI systems.

The growing demand for AI audits has driven the development of a heterogenous audit ecosystem. This ecosystem consists of private audit firms hired directly by AI developers and deployers and independent auditors such as academics and journalists who operate on behalf of the public [12]. But because there are so many approaches to AI auditing, it is sometimes difficult to know which audits best capture discrepancies and errors within the AI system, that is, which audits are of high quality [13]. AI auditors have varied standards and procedures for demonstrating that their audits are of comparable effectiveness and value [14]. Most auditors, as well as the systems and AI companies they audit, are located in minority world countries that have already enacted AI regulations [15]. Audits concerned with evaluating the social impacts of AI systems would need to attend to the concerns of not just AI developers, deployers, and regulators but also those of the many individuals and communities across the world adversely affected by these systems [16],[17]. If audits are meant to assure stakeholders of the benefits of AI, public trust in the audit ecosystem is critical.

Effective AI audits therefore require a framework that comprehensively evaluates whether an AI system aligns with stakeholder expectations and concerns regarding that system. Such a framework would also account for how AI's global impacts materialize in specific contexts. To lay the foundations for such a framework, this technical paper provides a brief overview of the AI audit ecosystem and articulates the strengths and limitations of current approaches to AI auditing[i]. Through this review of theory and practice, we identify the research consensus on effective AI audits. By applying these lessons, our proposal for a global AI auditing framework clarifies the essential features of an audit and provides guidance for translating important considerations in audit design and process into practice.

2. FEATURES OF AI AUDITS: WHAT THE AI AUDIT ECOSYSTEM TELLS US

AI auditors from a variety of backgrounds, including private firms, governmental institutions, academics, and civil society actors, have sought to meet the growing demand for assurance and accountability mechanisms (see Appendix A for examples of audit firms, governance efforts that require AI audits, and training or professionalization initiatives). Auditors in this diverse ecosystem demonstrate varied commitments, employ different approaches, and involve different stakeholders.

Audit approaches used by audit firms and discussed in academic literature are concerned with a few key questions about who conducts the audit, who the stakeholders are, what is audited, what conditions must be met to pass the audit, and what happens after the audit is concluded.

2.1. What is Audited and Why?

Organizations that commission audits, societal concerns around AI, and the regulations, standards, and/or cultural and ethical values motivating audits—their purpose—help define what is audited, that is, an audit object. AI systems' design, development, and deployment engage social structures and processes, making these systems sociotechnical in nature [12]. Audits might evaluate the organizational management or governance of an AI system, as required by the ISO 42001:2023 standard, or the system's outputs and outcomes, as stated in New York City's Local Law 144 [18],[19]. Specific features of a system's model(s) and data, such as model accuracy, data representativeness, or data provenance, can also be addressed in an AI audit. Which relationships between the technical AI system and its organizational and/or social environment are to be evaluated therefore defines the audit scope.

2.2. Who Conducts the Audit?

Internal auditors are employees of the organization who are assigned to review the AI system developed or deployed by that organization. There are two types of external auditors: third-party auditors, who are paid by the auditee, and independent auditors, who are *not* paid by

the auditee but may be funded through other means, such as grants, nonprofits, universities, or foundations [12]. Public agencies, such as the U.S. Government Accountability Office or the UK's Information Commissioner's Office, also conduct independent audits of AI systems [20],[21].

2.3. Who has a Stake in the Audit?

Various stakeholders—including but not limited to technical staff involved in developing the AI system, domain experts, and communities impacted by AI use—may be involved in the audit. Their input is considered an important part of the audit process, especially as evidence for the rigor and thoroughness of the audit's assessments. Stakeholders not directly involved with the development or deployment of a system, such as end-users and clinicians, can also be active participants in the audit process. International organizations such as UNESCO, the Global Partnership on AI (GPAI), or the Organisation for Economic Co-operation and Development (OECD) can also play a valuable role in connecting AI governance and auditing efforts across jurisdictions [22],[23].

2.4. What Conditions Must an AI System Meet to Pass the Audit?

AI systems are assessed against audit criteria based on collected evidence. Such criteria include performance metrics concerning bias and interpretability, developers' privacy practices, ethical considerations such as fairness, and measures of public safety [24],[25],[26]. Audits also rely on an expanded set of test results and documentary evidence that indicate how an AI system was designed, developed, and deployed, as well as how it is tested and monitored. Such audit evidence includes the results of bias tests, model cards, datasheets, design history files, stakeholder maps, and the results of (environmental and social) impact assessments and Failure Modes and Effects Analyses (FMEAs) [27],[28],[29].

2.5. How is an Audit Conducted?

Many AI audits are **conformity or compliance assessments**. They typically rely on conventional audit methodologies, tools, and instruments, such as checklists of evidence and

structured questionnaires, to collect audit evidence [19]. Based on this evidence, auditors assess whether a system's features comply with the audit criteria. Such audits can also discover any risks an AI system poses to organizations and people.

In contrast, **performance evaluation and harm discovery** audits directly test the outcomes of the AI system [12]. These evaluations, typically drawing on the methodologies of social scientific audit studies, investigate a system's outputs to identify problematic variations or inconsistencies, such as bias against individuals or groups. They can also evaluate whether the system's design, development, operation, and/or application may engender adverse societal and/or environmental impacts and harms.

These two broad types of audit methodologies may overlap. Conformity assessments, such as those laid out in the EU's AI Act or the Machine Learning for Health (ML4H) audit framework, can require performance evaluations [30],[31]. In such complementary approaches, audits test an AI developer's or deployer's claims about the performance of their AI systems or their compliance with AI governance efforts.

2.6. What Happens After an Audit?

After an audit is completed, legislative, regulatory, standards, and certification regimes such as ISO 42001 and New York City's Local Law 144 can require post-audit activities. They may include a follow-up audit at some set frequency and notifying the relevant stakeholders, such as the public, of the audit results [32],[33].

This section's survey of the audit ecosystem has identified the important features of any AI audit: the audit **scope(s)**; who the **auditor** is; which **stakeholders** might be involved in the audit; against what **criteria** and with what **evidence** the AI system is assessed; what **methodology** is used to conduct the audit; and what the **post-audit activities** are.

3. CHALLENGES WITHIN THE EXISTING AI AUDIT ECOSYSTEM

The AI audit ecosystem still faces several challenges to its efficacy and trustworthiness, especially in addressing the disparate global impacts of AI systems. These challenges stem from:

- inadequacies in audit design and conduct;
- auditors' limited access to AI systems and documentation;
- lack of clear bases for establishing auditor independence and qualifications;
- limited accessibility of audit reports.

When the design and conduct of an AI audit follows the limited requirements set by developers, deployers, or regulators—typically in wealthy countries—important societal concerns around AI might be missed [34]. For example, few auditing frameworks recommend that audit criteria account for an AI developer's adherence to labor laws. Other global concerns, such as violations of human rights or Indigenous data sovereignty, are often missing from audits [35],[36],[37]. Further, audits have largely determined whether an AI developer or deployer has conducted harm discovery, such as through impact assessments or bias testing. Rarely do they assess whether the auditee has conducted harm mitigation, such as algorithm and data debiasing. Put simply, audits often determine whether an AI system's impacts, risks, and potential harms have been documented and (where relevant) disclosed, but not whether concerns related to those effects have been addressed.

Another challenge to the quality and trustworthiness of AI audits is auditor access to adequate evidence about a system's design, development, deployment, and operations. This access is usually limited by the terms of the auditor-auditee relationship through mechanisms such as bilateral information-sharing agreements, which can lead to information asymmetries. Auditors conducting conformity assessments, especially on behalf of the AI developer or deployer, typically rely on documentation provided by the auditee. Auditors who are independent of the developer or deployer might not have adequate access to system

documentation, which poses significant difficulties to evaluating AI systems along their life cycle and supply chain. Geo-economic disparities may also impede auditors' activities in majority world countries by limiting their access to evidence [34],[39]. Further, auditors might not have access to the system to test it themselves [40].

Another key challenge to audit trustworthiness and quality is establishing auditors' independence and qualifications. Third-party auditors, typically audit firms, are primarily accountable to the management and shareholders of the company being audited—the very parties who pay these auditors. This can result in a perceived conflict of interest—given the financial relationship, has the audit been conducted in a way that can meaningfully reveal problems with an AI system and its management? Traditionally, in financial or operational audits, an auditor's independence is typically established through their conformity with legal and professional standards such as those established by the USA's Public Company Accounting Oversight Board (PCAOB) or the Institute of Internal Auditors (IIA) [41],[42].

Few such standards exist for AI auditors, although nonprofit organizations such as ForHumanity have developed voluntary ones [43]. However, it remains difficult to verify auditor adherence to existing standards. Another open question is how an auditor can demonstrate that they are qualified to conduct AI audits [44],[45],[46]. There are currently few professional training programs for AI auditors. Some professional associations of auditors and accountants, such as the Institute for Internal Auditors and the ISACA, offer guidelines for applying their existing auditing frameworks to AI systems. The high costs of this training and certification and its lack of linguistic diversity are barriers to global participation[ii], constraining the effectiveness of the audit ecosystem.

Comprehensive audit reports are a key element of a high-quality audit [48]. Which elements of the audit design, process, and procedure are documented in the report can limit audit replicability and stakeholder ability to verify its results. Further, audit firms' reports are often only available to the party that commissioned the audit, although some auditors and auditees have publicized details about audits through academic and news publications [49].

Stakeholders outside the auditor-auditee relationship usually only have access to summaries of audit reports, which may obscure important information about the development, deployment, and operations of AI systems.

Given the practical challenges AI audits still face, it is instructive to examine the opportunities and limitations of a current audit regime. In the following section, we illustrate how well the design, processes, and procedures of a pioneering AI auditing law correspond to the features of an effective audit and what the law's limitations reveal about how audits are conceptualized.

4. CASE STUDY: BIAS AUDIT PER NEW YORK CITY'S LOCAL LAW 144

In this case study, we examine a bias audit of an AI-based Automated Employment Decision Tool (AEDT) [50]. The audit was conducted by a private firm to comply with New York City's (NYC) Local Law 144 (LL144) and its subsequent 2023 rule [19]. LL144 has required audits when employers and employment agencies in NYC use an AEDT, regardless of the level of risk posed by a particular system or the size of the AEDT deployer. This study elucidates the auditor's role and how the audit defines its scope, criteria, evidentiary requirements, methodology, and post-audit activities. Based on this examination, we recount what lessons LL144 offers for an auditing framework suited to AI's disparate global impacts.

AEDTs can be used at any stage of the recruitment and hiring process. Although jobseekers and employees may not directly interact with an AEDT, the system relies on data provided by those persons or by employers to, say, predict a jobseeker's capabilities or future performance. AEDTs are typically used to complement—rather than replace—human processes. This AEDT provides insights that aim to support HR decision-making without replacing human judgment and therefore, in the language of the rule, "substantially assist[s]" with hiring [19].

4.1. Audit Scope

The purpose of this particular bias audit was to show compliance with LL144. The object of the audit was the AEDT's bias, as defined and required in LL144. The "auditee" developer Eightfold's AEDT utilizes machine learning to generate a "match score" that describes the fit of job candidates to a position. This score, ranging from 0 to 5 in 0.5 increments, assesses the compatibility of a candidate with a job. A candidate's score is not fixed; it varies with different job pairings and updates candidate-related or job-related information. The AEDT is intended to help rank candidates for job positions and allows candidates to find jobs matching their skills on Eightfold-powered job pages or portals.

LL144 encodes the audit object in law, which ensures that any bias audit must—at a minimum—address an AEDT's disparate impact on various groups of people across gender and ethnicity. Such a legal definition can act as a bulwark against "audit-washing"; it can ensure that problematic aspects of the AI system—like potential bias against racialized minorities—are uniformly audited [51]. By clearly defining the audit scope, LL144 ensures that any audit conducted according to this regime addresses an AEDT's bias. Such clarity also eases AI deployers' compliance efforts.

The clarity and definitions provided by LL144 are very valuable. We note that bias audits conducted according to LL144 are limited, though, to evaluating the measurable outcomes of an AI system, specifically, biased hiring decisions. The law does not take into account impacts that emerge from how an AEDT is used in the hiring pipeline. For example, if jobseekers are unable to opt out of interacting with an AEDT, they are forced to either allow the system to collect their data or forfeit the opportunity. Furthermore, centering an AI audit on the system's outputs, rather than accounting for its organizational context also risks reifying the conditions under which the organization might deploy that system. The underlying need for an AEDT typically comes from hiring departments being under-resourced—a problem that cannot be addressed by a bias audit. LL144's defined audit scope may not, therefore, account for public concerns related to a system's broader sociotechnical context [52],[53].

LL144 clarifies the audit scope with the aim of having AEDT deployers demonstrate that their system does not produce biased hiring decisions. Given how AEDTs are used in the hiring pipeline, however, it is important for audits to consider how the system's societal impacts go beyond its direct outputs.

4.2. Auditor

We chose the bias audit conducted by BABL AI, a specialist AI audit firm (referred to in this section as "the auditor"), to keep continuity with organizations described in Appendix A [iii]. The auditor must meet three responsibilities: 1) obtain reasonable assurances from the auditee that their statements were truthful; 2) determine whether these statements were

sufficient evidence to satisfy the audit criteria; and 3) issue a full audit report including their opinion on the audit results. The auditor also has the responsibility of accurately evaluating the AEDT, for which they are accountable to the auditee.

LL144 outlines auditor independence as an important condition for conducting bias audits, as it helps assure stakeholders that the auditor will honestly state their findings. The auditor in this case study meets LL144's conditions for independence. It also conforms to definitions of independence and objectivity laid out by ForHumanity's code of ethics for AI auditors and with the Sarbanes-Oxley Act, which is used to regulate financial audits in the USA [43]. They were not involved in the development, distribution, or use of the auditee's AEDT. The auditor declared that their contractual and financial relationship with the auditee was limited to services rendered for assessing LL144 compliance and had no bearing on the final audit opinion.

LL144 does not, however, preclude an audit firm from being hired by the auditee. In such instances, auditor independence is at least somewhat contingent on the financial relationship between the audit firm and the auditee. How rigorously an auditor assesses an AI system is bounded by the minimum degree of scrutiny that potential auditees want and the maximum that they will accept: an auditor is less likely to issue an unfavorable audit opinion if it would deter the auditee from rehiring them [54]. Audit firms might self-declare their independence—as BABL AI does—but the absence of a legally enforced definition and strict auditing standards can raise questions around verifying such claims.

The audit scope defined by LL144 also indicates that an auditor might require expertise in disparate impact analyses and USA employment law to conduct a bias audit. However, LL144 does not specify how an auditor might demonstrate their qualifications for AI auditing. Specialized AI auditing firms like BABL AI therefore differentiate themselves from traditional audit firms like Deloitte through their experience and expertise in AI audits. BABL AI's status in the ecosystem is bolstered by its own provision of AI auditor training and professional certifications and by its role as a founding member of the International Association of

Algorithmic Auditors (IAAA). Existing private certifications like those provided by BABL AI are, however, expensive and limited to English speakers. This constrains auditors across the world with fewer resources from accessing training and professionalization, potentially shutting them out of the audit ecosystem.

LL144 clearly outlines an auditor's responsibilities when conducting a bias audit and makes auditor independence a cornerstone of its auditing regime. However, the law leaves open the question of how auditors can demonstrate their independence and qualifications when they are under practical constraints.

4.3. Audit Criteria

The bias audit described here evaluates the AEDT against two key metrics established in LL144: the selection rate (the proportion of applicants advancing from one hiring stage to the next) and the impact ratio (the selection rate for a specific category divided by the selection rate for the most selected category). Categories are defined by race/ethnicity and gender classifications set by the U.S. Equal Employment Opportunity Commission (EEOC), as identified in the submitted job applications. This process calculates the disparate impact of the AEDT's outputs on these different demographic groups. The audit included calculations for binary sex categories (male and female), race/ethnicity categories (for example, Hispanic or Latino, White, Black or African American), and intersectional categories (such as Hispanic or Latino males and Black or African American females). In addition, the auditors set criteria for assessing Eightfold's organizational governance.

Setting a specific bias metric like the impact ratio as the audit criterion may not, however, be well suited to globally diverse audit contexts. The disparate impact test mandated by LL144 may fail to detect bias in an AEDT if the historical or test data contains certain distributions of characteristics [55]. There are also ongoing debates about the relative merits of different bias and fairness metrics, which are both useful for uncovering a particular form of algorithmic bias [56]. Given the sociotechnical character of AI systems, a useful bias metric would align with the social context in which the system operates. For example, in the USA, where Black

communities are often targeted by predatory lending practices, auditing an algorithm for disparate treatment may be more appropriate than focusing on disparate impact [57],[58]. Additionally, race or ethnicity may not always be the most appropriate social dimension along which fairness can be assessed. In many contexts across the world, gender, religion, caste, ability, or language might be more relevant vectors of bias [15],[59].

Audits often also aim to evaluate the impacts of an AI system for which bias metrics, designed for its measurable outputs, might be ill-suited. Crucially, such quantitative criteria cannot capture system impacts that are difficult to directly measure but are socially consequential. For example, public concerns around AEDT-supported hiring often point to the dearth of mechanisms that can be used to challenge the system's decisions [33]. Assessing an AI developer's accountability mechanisms, then, requires criteria that go beyond bias.

Notably, LL144 does not impose a minimum threshold that the calculated impact ratio must meet to "pass" the audit; it is sufficient for the auditee to demonstrate that they have calculated these metrics. While such documentation helps enforcement agencies such as the EEOC hold AI deployers accountable, it does not preempt biased decision-making. In contrast, some independent audits have assessed whether a system's decisions comply with the four-fifths rule, which requires a minimum impact ratio of 0.8 across demographic groups—a common standard for evaluating employment parity in the USA. [49].

LL144 and its corresponding rule clearly outline the criteria against which an AEDT must be assessed. Assessing whether a system has disparate impacts on demographic groups—measured using the impact ratio for AEDTs—is commonly accepted as a useful test for detecting employment discrimination. It is clear to both auditors and auditees what would constitute compliance with the law, helping ensure that audits of different AEDTs are based on shared grounds of assessment. However, bias metrics such as those mandated by LL144 may not be useful in every audit. Social and cultural expectations of AI systems—including fundamental questions such as what counts as a problem with AI—often vary according to context, requiring audit criteria that can account for these differences.

4.4. Audit Evidence

The evidence used in the bias audit, namely the results of disparate impact testing and the auditee's documentation of historical data used in the AEDT, draws directly from the LL144 rule. The auditor in this case study also collected evidence about Eightfold's organizational governance, such as documentation indicating the chains of responsibility for the AEDT.

Similar to how LL144's definitions of audit scope and criteria are beneficial, the clear definition of what counts as evidence helps ensure consistency across audits. Additionally, LL144 allows auditors to use testing data that employers and employment agencies in NYC besides the auditee have collected. This enables audits to be conducted even when an AEDT deployer cannot provide historical data about the system, ensuring that it can be tested even before it is put into widespread use.

The evidentiary requirements for LL144's bias audits, however, risk making auditors reliant on information and datasets provided by the AEDT developer themselves. Since bias test results provided by the AI developer or deployer themselves are acceptable audit evidence according to LL144, such evidence need not be produced by the auditor. In addition, LL144 does not require the auditee to record why certain data is missing from its historical dataset or to demonstrate how they have accounted for missing data [60]. Benchmark datasets that do not adequately account for underrepresented social groups risk reproducing the biases and sociocultural disparities enacted by AI systems [61]. Overreliance on auditee documentation also becomes problematic when auditors are facing challenges regarding accessing data from AEDT vendors. These entities, who are the prescribed auditees in LL144, may not be able to provide all the information required for an audit. This is a particular roadblock when the AI vendor is in a different jurisdiction than the system developer [62]. Limiting evidence to what can be collected from an AI developer or deployer in a local context, then, may be insufficient for an effective audit.

LL144 and its corresponding rule clearly outline the evidence with which an AEDT must be assessed. However, that very prescription of audit evidence can constrain what auditors and auditees record about an AI system.

4.5. Audit Methodology

The auditor's bias audit methodology was threefold: 1) scoping the auditee's algorithm to properly contextualize the documentary evidence (that is, bias test results and information about job applicants); 2) evaluating and verifying the submitted documentation to demonstrate satisfaction of the audit criteria; and 3) providing certification if the bias audit was passed, whereby the auditor would certify the audited algorithm as LL144-compliant.

The evaluation and verification part of the auditor's methodology required them to check whether the auditee had properly calculated and recorded the selection rates and impact ratios for each demographic group in the gender, race/ethnicity, and intersectional categories. In accordance with LL144's rule, the audit report also included the number of individuals who were not included in the bias test's calculations due to missing data such as insufficient demographic information.

While LL144 does not prescribe any particular way to conduct a bias audit, its specification of audit scope, criteria, and evidence can help auditors develop an appropriate methodology. This flexibility is useful when auditing AEDTs in different application domains. The auditor can also apply methods and tools based on what evidence they have access to, what documentation the auditee provides, and what performance tests are feasible.

LL144 does not, however, specify what makes an effective audit methodology. It therefore leaves open the risk of inadequate audits being conducted. Notably, the law does not require the auditor to test an AEDT themselves; the auditor instead relies on the results of bias tests conducted by the AI deployer themselves. This allows the deployer to obscure system outcomes or behaviors that can lead to biased decision-making. LL144 also does not require auditors to calculate an AEDT's impact ratio or selection rate for people missing from the historical or test datasets. Conducting bias tests that do not properly account for missing

data might lead the audit to ignore an AEDT's outcomes for marginalized social groups that might be underrepresented in the existing data [60]. Missing data also makes it impossible to account for those groups that chose not to self-report their personal information.

Furthermore, LL144 does not offer ways to accommodate technical developments in AEDTs that make the calculation of the outline criteria difficult. Although LL144 is tailored to predictive models with discrete and anticipatable outputs (that is, "closed-output" systems), it is less suited to "open-output" systems like large language models (LLMs), whose outcomes and potential uses are less predictable. LLMs and other generative AI systems are difficult to audit because they can be used unpredictably [63],[64]. Despite the evolving nature of AI systems, an effective audit methodology should retain some fundamental characteristics that demonstrate its quality [65].

LL144 does not prescribe an audit methodology, allowing auditors to adapt to the demands of a specific audit context. At the same time, the law does not clarify what makes for a good audit methodology. Without clarifying which methods and tools are suitable for what kinds of assessments, there is a risk of conducting ineffective audits that may miss important AI impacts.

4.6. Post-Audit Activities

LL144 requires that the auditor make publicly available a summary of the audit results. In addition, employers and employment agencies in NYC are required to provide "transparency notices" to employees and job candidates disclosing their use of AEDTs and the corresponding bias audit results. LL144 also requires that subsequent bias audits must be conducted annually to ensure compliance and continued transparency by the AEDT user. In accordance with these requirements, the auditee in this case made the summary of their audit report—conducted on the 3rd of April 2024—publicly available. This summary can be accessed through the auditee's website and search engines [50].

LL144 marks a significant regulatory push towards transparency around the use of AI systems. The public disclosure of audit results can help stakeholders identify the potential harms of a system and, if necessary, move an enforcement agency such as the U.S. Equal Employment Opportunity Commission (EEOC) to act [62]. This also puts pressure on employers using AEDTs to adhere to anti-discrimination laws or face litigation. Through transparency notices, jobseekers can also be made aware of how their data might be used by a prospective employer. They can then decide on further actions, such as continuing to share their data or contesting AEDT use.

However, it is often difficult to verify regulatory compliance with LL144. It does not prescribe any mechanism for disclosing the audit report or transparency notices, making it unclear how those post-audit disclosures might function in practice. Furthermore, it is impossible to determine whether the lack of such disclosure means that an employer does not use an AEDT or whether the employer has failed to disclose their use of an AEDT. LL144, then, often leads to a state of "null compliance", where it is unclear whether "the absence of evidence of compliance", such as an audit report or transparency notice from an NYC employer, necessarily means that the auditee is not compliant with the law [66].

Another limitation of LL144—owing, presumably, to jurisdictional differences— is that it does not link the audit results to action by the EEOC or other relevant enforcement agencies [62]. This, combined with the inconsistency of post-audit disclosure, means that there is no clear mechanism for the audit results to be tied to accountability structures such as courts. LL144 therefore seems to impose additional costs on employers—annual audits, for example, become a recurring expense— without meaningfully ensuring public accountability of AEDTs. Audit regimes such as those set up by LL144 can therefore best be described as transparency and disclosure regimes—meaning that an AEDT's disparate impacts might be identified without being remedied [67].

Public disclosure of audit reports and post-audit user notifications are a vital step towards inculcating trust in the audit ecosystem: it demonstrates that there is nothing about an AI

system's operations and outcomes that its developer or deployer wants to obscure. Consistent disclosures that can be used in oversight and accountability mechanisms also help demonstrate that AI developers and deployers are making necessary improvements to their system.

As this case study demonstrates, an audit's design, processes, and procedures hew closely to legislative or regulatory requirements, such as those laid out in LL144. However, this case study also reiterates four key lessons for conducting effective audits and developing trust in the audit ecosystem:

- An auditor's independence and qualifications are key to an effective audit.
- An AI system's contexts influence how the audit is designed and conducted.
- An auditor conducting their own performance evaluations of a system and having access to model and data documentation benefits the audit's effectiveness.
- Public disclosure of audit results that can be linked to accountability processes, including regulatory enforcement, fosters trust in audits.

These lessons offer vital considerations when developing a general framework for effective audits. As AI systems proliferate globally, audits of comparable quality across different contexts become crucial.

5. WHY A GLOBAL AI AUDITING FRAMEWORK?

It is often unclear how high-level guidance and principles found in AI auditing research and AI governance efforts can be translated into practice, especially given diverse sociocultural contexts [68],[69]. The lack of a common framework makes it further difficult to compare the effectiveness of heterogeneous auditing approaches. A global AI auditing framework therefore is needed to bridge the gaps between multidisciplinary research on auditing and practical auditing process and procedure.

A global auditing framework for AI systems coalesces expert consensus in a systematic format that accounts for the needs of industry, regulatory actors, and the public [45],[70]. It establishes a common basis for comprehensive and comparable assessments of an AI system by providing guidance on how to determine:

- the scope of an audit;
- the role of an auditor, including their qualifications and independence;
- the context of an audit;
- stakeholders' involvement;
- the various criteria that should be part of an audit's assessments;
- what documentation and artifacts are used as audit evidence;
- which methodologies are appropriate;
- what happens after the audit.

Our proposed framework offers suitable and useful paths to designing and conducting an effective audit without placing undue constraints on it [71]. To create common ground, the framework uses a shared vocabulary informed by expertise from academia, civil society, governance structures, and global industry. It helps shape audits specific to a local context

that also take into account AI's global impacts through meaningful multistakeholder inclusion throughout the audit [72],[73],[74].

The framework provides the scaffolding steps for any AI audit, with additional considerations for each audit feature that should be applied depending on the audit's context. Not every consideration in this framework applies to every audit. This framework can be used to demonstrate auditors' commitment to good practices while allowing them to flexibly apply their discretion and judgment when conducting the audit. Following this framework would also allow auditees to preemptively demonstrate responsible and trustworthy AI development and deployment, potentially relieving their compliance burdens.

6. THE GLOBAL AI AUDITING FRAMEWORK

This framework recommends that the following be done by the auditor and/or auditee in conjunction with relevant stakeholders. These tasks are not necessarily sequential, as some may happen in parallel or need to be revisited as others are carried out.

- Determine the **scope** of the audit according to the AI system layer and stage(s) of the AI life cycle and supply chain being audited. The scope should define the **object(s)** according to the **purpose** of the audit.
- Clarify the **auditor's** role, identifying the stakeholders to whom the auditor (or audit team) is accountable and the relationships that may exist between them.
- Map the **stakeholders** relevant to the audit's **contexts**.
- Develop structured participatory methods for ensuring **stakeholder engagement** throughout the audit process.
- Establish the **criteria** used to determine an AI system's regulatory compliance, acceptable technical performance, and/or minimization of societal and cultural impacts (including on labor and the environment).
- Identify what **evidence** is required to properly evaluate an audit object against the established criteria and how that evidence can be collected.
- Develop the audit **methodology** best suited to assess an audit object against the established criteria.
- Check whether the audit meets the expectations of the stakeholders involved throughout the audit, **revisiting** aspects that they challenge.
- **Conduct** the audit, employing system performance tests, documentary analysis, and/or gather evidence from stakeholders, along with any other requisite evaluations.

- Develop the **audit report** and submit it to the relevant stakeholders. If possible, make the audit report publicly available.
- Establish the conditions for **follow-up audits**.

Given the sheer variability in audit design and practice, it is infeasible to prescribe specific audit processes and procedures in this framework. Instead, we provide general recommendations and directions for designing and conducting an effective audit that is attentive to AI's disparate global impacts.

The framework is subject to additional considerations that clarify audit design and processes. Any single AI audit cannot—and should not need to—apply to every aspect of this framework. Choosing which of these recommendations to apply depends on the audit's context, the auditor's resources and capacity, and which aspects of an AI system are of interest to stakeholders.

Lastly, these recommendations necessitate further auditing research and institutional capacity-building. The means to operationalize some of this framework's recommendations, such as documentary analyses that require data access by AI developers, may still be unavailable to auditors.

The following subsections describe in detail our recommendations for developing audit features—**scope, auditor role, context and stakeholders, criteria and evidence, methodology,** and **post-audit activities**—for effective audits.

6.1. Audit Scope

The scope of the audit is defined in terms of its purpose and object(s). Recommendations to determine the scope are as follows:

a. **If an audit seeks primarily to evaluate an AI system's performance, the audit object(s) should be selected from the relevant performance characteristics, such as accuracy, reliability, fairness, transparency, interpretability and explainability, and robustness and safety.**

Explanation: Auditing system performance offers multiple ways to assess whether an AI system meets technical and societal expectations. Evaluating an AI system's:

- **accuracy** reveals how it aligns with 'ground truths' that can be drawn from social norms and ethical commitments[iv] [75],[76].
- **fairness** checks whether the outputs of an AI system are unfairly biased against certain individuals and/or groups of people [77].
- **transparency** documents what has been disclosed about the constituent parts of an AI system—its model, data and how it interacts with social systems [78].
- **interpretability and explainability** determines whether stakeholders can understand why a system produces the outputs that it does, and what those outputs mean [79],[80].
- **robustness** evaluates how well a system performs when it is subject to (potentially malicious) interference, unexpected inputs, or other adversarial conditions [40].

b. **If an audit seeks primarily to evaluate regulatory or standards compliance, the audit object(s) should be defined according to the appropriate regulations or standards.**

c. **If an audit seeks primarily to assess the risks an AI system poses to an organization, the audit object(s) should be defined in alignment with organizational policies and standards.**

d. **If an audit responds primarily to societal concerns around AI, the audit object should be defined according to social, cultural, and ethical considerations.**

Explanation for b.–d.: The regulation or standard against which the system is being assessed differs depending on the jurisdiction or the country in which the audit is conducted, which can vary over the AI system's life cycle. Organizational policies cover an organization's

mission, goals, and values. AI's impacts across multiple domains of human life have meant that, increasingly, algorithmic bias and AI's social, economic, cultural, and environmental impacts have also come under consideration [81]. The audit object(s) should therefore be defined accordingly.

e. **The auditor should establish which aspect(s) of the AI system—the model(s), data, (organizational or regulatory) governance, and/or (direct or indirect) outcomes—matter while defining the audit object(s).**

Explanation: The aspects of the AI system being audited affect how the audit object(s) are defined:

- The objects of **model** audits vary depending on the model type (foundation, bespoke, fine-tuned) and model features, such as the source code or model weights, that can be evaluated [82].
- When auditing the training, testing, and/or evaluation **data** used in a system, the object can be a (closed) commercial or open dataset, with closed datasets limiting what aspects of the dataset can be audited [83]. Many of the datasets that have the most impact, such as those used by OpenAI to train later models in their GPT series, cannot be directly audited. Such limitations matter when auditing data provenance, quality, and representativeness. In some such cases, it may be possible to infer the features of a data-related audit object through black-box tests of the model [84],[85].
- Audits of an AI system's **governance structures**—including oversight and monitoring processes—must define the object in terms of the relevant policies and regulations [18],[28]. Often, an audit of an AI developer's or deployer's conformity with organizational policy also assesses regulatory compliance.
- An audit of a system's **outcomes** typically defines the object in terms of the system's social and environmental impacts. Audit objects can include bias or harmful content in the system's direct outputs. Audits can also evaluate the indirect impacts on

democratic processes, labor, and the environment that emerge in the communities and locations where a system is developed, deployed, or operated [86],[87]. Some audit objects, such as system fairness, can account for both a system's direct and indirect outcomes [81].

f. **The auditor should establish which stage(s) of the AI life cycle—design, development, and/or deployment—and which parts of the supply chain—labor and material—matters while defining the audit object(s).**

Explanation: Different parts of an AI system can be audited at different stages of its life cycle and the AI supply chain. Audits conducted during the design or development stage, for example, can assess whether the model developed for the system is appropriate for its intended use [27]. Audits conducted during the deployment stage, in contrast, typically focus on an AI system's outputs [65]. Audits of the labor supply chain can consider different forms of work, such as data cleaning or maintenance "patchwork" that ensures the smooth functioning of automated systems [88],[89]. They can also define the audit object in terms of socioeconomic conditions, such as fair compensation [90]. Audits of the material AI supply chain can assess the AI developer's or deployer's hardware and software sourcing procedures, such as whether the components have been purchased from a certified supplier and the environmental sustainability of the supply chain [18].

As this section makes clear, the object of AI audits is not limited to the motivations of AI developers or deployers. Precise scope definition is crucial for developing targeted criteria and methodologies [12].

6.2. Auditor Role

The auditor's relationship with the auditee is an important factor in the audit. Our recommendations for clarifying the auditor's relationships and capabilities are as follows:

a. **For the defined audit scope, the auditee or concerned stakeholders should determine whether internal or external (third-party or independent) auditors would be most appropriate.**

b. **An auditor's obligations and accountability to the auditee and stakeholders should be clarified.**

c. **How an auditor is being paid or funded and by whom should be declared.**

d. **Auditors should verifiably adhere to standards of independence and be accountable to oversight mechanisms. They should note challenges they encountered during the audit.**

Explanation for a.–d.: Internal auditors and third-party auditors are typically accountable to the auditee's management and/or shareholders, who are concerned with adherence to organizational policy and risk tolerance [91]. Third-party auditors also balance their commitments to the auditee with regulators' accountability requirements—including meeting independence standards. Who they are accountable to and who funds them, however, can vary. For example, in some audits, the intended beneficiaries of a certain welfare scheme are paid by the government to evaluate whether public funds have been properly used in that scheme. These community auditors remain primarily accountable to welfare beneficiaries rather than to the audited government entity [92],[93]. Independent auditors, such as journalists and academics, often do not have a contractual relationship with the auditee and seek to hold AI developers and deployers accountable on behalf of the public.

To show auditor independence given financial dependencies, internal and third-party auditors can state their adherence to standards of independence set by regulators or civil society actors [44]. These auditors, however, are often subject to financial pressures, including the risk of the auditee choosing to hire another firm more likely to issue a favorable audit opinion [94],[95]. An audit firm might also be hired according to political or corporate favoritism, especially when standards for independence are difficult to enforce [96]. An economically or politically powerful AI developer or deployer might also pressure auditors to issue a positive opinion—despite stakeholder concerns—to preserve the auditee's interests, as has been the case with "social accountability" or "ethical" audits of supply chains and

forestry [90]. In contrast, independent auditors are not financially beholden to the auditee. They may instead rely on funding from government, research grants, clients, or their employer, which they should disclose. Academics and journalists should adhere to professional standards to demonstrate independence. It is important to note, however, that non-auditee funding for independent audits remains significantly limited.

e. **Auditors should demonstrate why and how they are qualified to perform the audit.**

Explanation: Having rigorous qualifications for auditors allows them to demonstrate their competency to various stakeholders and helps engender trust in the audit process [97]. Currently, the limited professional qualification and training programs in the AI audit ecosystem that exist are concentrated in global minority countries and mainly offered in English. Leaving credentialing power to private initiatives, therefore, can limit who counts as a legitimate auditor [85].

Auditors demonstrate their competencies through multiple means. They might be able to translate their expertise in other forms of auditing or accounting, such as operational and financial audits, to AI audits. Privacy- and cybersecurity-related qualifications can help auditors demonstrate that they are well equipped to handle sensitive data and proprietary systems. Domain expertise and/or academic credentials can allow auditors to demonstrate their qualifications to evaluate specific aspects of AI systems, such as their technical performance or socioeconomic impacts. Experience with community advocacy and public interest investigations can help independent auditors, such as activists and journalists, show that they can work with end-users or communities impacted by AI systems to conduct effective audits.

f. **Where necessary, audits should be conducted by a multidisciplinary and multi-domain team with knowledge of the particular context.**

Explanation: Given the range of AI's impacts and life cycle, a single auditor may not have the necessary expertise to assess a particular system. AI's globally dispersed supply chains and

application domains means that auditors across geographies might need to pool their findings to issue the final audit opinion. Furthermore, a diverse audit team can holistically assess the system and conduct high-quality audits [86]. Audit frameworks for medical algorithms, for example, involve both test engineers and domain experts such as clinicians [31],[100],[101]. Auditing AEDTs benefits from involving experts across psychology, law, and computer science [52]. An audit team with a diverse set of experiences and viewpoints can also better account for stakeholders' varied needs.

g. **The auditor should receive assurance of protections, such as safe harbor, when conducting independent audits.**

Explanation: AI developers or deployers can use the absence of legal protections to deter independent audits. Such protections are especially important when auditors use adversarial methods such as red teaming, in which an auditor can reveal major deficiencies in an AI system by simulating an attack. These methods risk adversely impacting a system's functionality [102]. Safe harbor therefore allows auditors to collect accurate data on an AI system's real-world operations without running inordinate legal risks [65].

6.3. Context and Stakeholders

To account for an AI system's local impacts across its life cycle, auditors should identify an AI system's contexts and develop well-structured processes and mechanisms for stakeholder engagement. Our recommendations for doing this are:

a. **The institutional, jurisdictional, economic, cultural, social, and/or domains with which an AI system interacts should be identified.**

Explanation: An AI system's design, development, and deployment depend in several ways on those processes' contexts, which also affect how an audit is carried out. For example:

- The same generic models might be used for financial services and in clinical diagnosis. Even though they can be fine-tuned using domain-specific data, the context in which they are deployed presents unique challenges and opportunities. An AI system used

by doctors can be subject to normal medical checks and balances throughout the diagnostic process, but the same kind of chatbot offered directly to the public by a bank with no supervision could be a disaster, for example if it offered customers unhelpful or even erroneous financial advice [31],[103].

- While facial recognition can be used in some jurisdictions, it may be banned in others [104].
- Economic incentives, such as the availability of cheap data labor or electricity, influence where the AI supply chain is located [6].
- AI systems can be biased against historically overpoliced groups, but also what counts as an overpoliced group varies by country [105]. Specific social and cultural norms affect what gets modeled by AI systems and how.

Identifying the context of the AI system—and therefore the audit—clarifies what can and cannot be audited and under what conditions, what information about a system can be collected, and what methods and tests can apply to a particular system. Certain contexts, such as the jurisdiction in which a system is deployed, can also affect what system information the auditor can reasonably access. It might be difficult for an AI deployer in a majority world country to provide an auditor with documentary evidence, such as data provenance records, that are in the hands of an AI developer located in another country. Clarifying such conditions also helps ensure that audits do not overburden under-resourced auditors and auditees [62].

b. **The potential harms of an AI system in an interaction domain should be identified.**

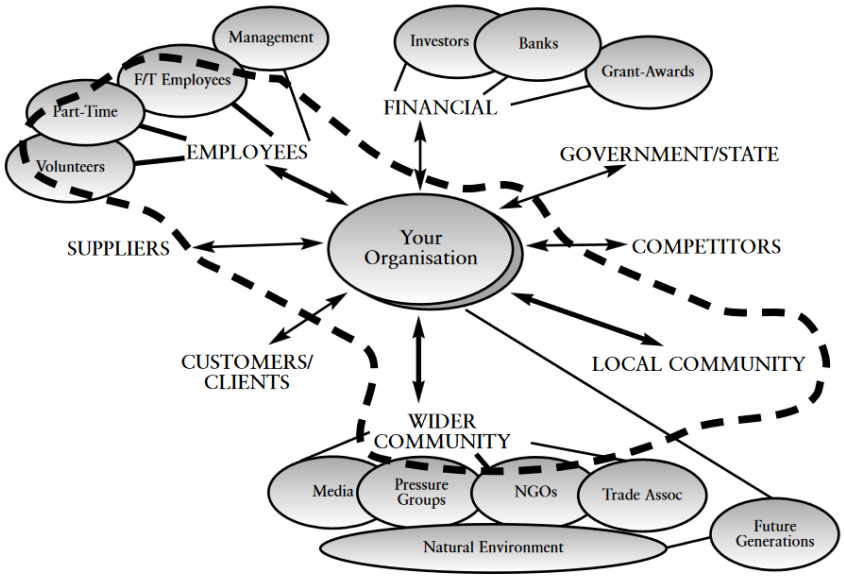
Explanation: The kind and degree of harm posed by a system depends on domain: computer vision used in self-driving cars carries different dangers than when used in marketing.

c. **Map those responsible for the design, development, deployment, oversight, and monitoring of an AI system, those directly or indirectly impacted by it, such as**

users and workers, and other individuals who can offer specific insights into the system. This mapping should also identify relationships between stakeholders and the system and among each other.

Explanation: It is only beneficial to audit an AI system if it is of interest to a well-defined set of stakeholders. Stakeholders include those who want to understand the system's risks and those who are directly or indirectly affected by the system [27]. They are not limited to actors such as shareholders invested in the financial, reputational, or operational health of the AI developer or deployer. Members of the public need to be engaged in the audit since their quality of life along dimensions such as access to public goods and economic opportunities can be materially affected by AI systems. Other stakeholders such as civil society actors, grassroots activists, domain experts, and legal experts pay close attention to AI's impacts in areas such as human rights and community advocacy that developers and deployers are not typically concerned with. Stakeholders can also include nonprofit organizations like data cooperatives and platform cooperatives whose goals align with an audit's larger concerns, such as democratizing the digital realm and advocating for the responsible and ethical use of AI. Properly defining the stakeholder map helps the auditor identify "all the potential accountabilities" implicated by an AI system; the map aims to show a holistic view of AI's sociotechnical relations [106]. It doubles as a valuable part of the audit trail by recording which stakeholders were engaged during the audit.

Figure 1. Mapping stakeholders for social accountability audits



Source: (reproduced [106, p. 260]).

d. **Stakeholder engagement should be conducted through participatory methods that account for barriers to participation. These methods should include appropriate tools to individually or collectively gather stakeholder input; aggregate and translate input into actionable considerations for the audit; explain to stakeholders how their input was used; and provide compensation (monetary or in-kind) for their time.**

Explanation: Treating stakeholder engagement as an iterative and collaborative process demonstrates that stakeholder input is valued and actively shapes key audit features [107]. Incorporating and valuing the voices of historically marginalized individuals and communities also requires auditors to meet stakeholders "where they are", that is, respecting individuals and communities' cultural differences, resource constraints, and linguistic limitations [108],[109].

Choosing the right mechanism for engaging with stakeholders requires the auditor to establish what input they aim to gather and what resources they have access to. Based on

this understanding, auditors can use interviews, focus groups, (semi)structured questionnaires, codesign workshops, community-oriented reflective tools, and citizen forums. Participatory frameworks such as WeBuildAI offer cues for developing appropriate methods and tools for stakeholder engagement, while communities such as Queer in AI can act as models for continuous and reciprocal auditor-stakeholder relationships [110],[111]. At a minimum, these participatory mechanisms should result in meaningful engagement: stakeholders should get to know that their input matters and will be used in concrete ways [107].

6.4. Criteria

Audit criteria should effectively capture stakeholder concerns about the risks of an AI system. When developing audit criteria, consider the following:

a. **Audit criteria should be precise, observable, well aligned with the audit object(s), and designed to assess risk or harm mitigation.**

b. **Develop criteria through meaningful engagement with stakeholders and by incorporating public knowledge.**

c. **Establish clear minimum conditions or thresholds for compliance with audit criteria.**

Explanation: Having well-defined audit criteria enables auditors to make reasonable judgments about an AI system. Additionally, selecting criteria based on what stakeholders consider acceptable system behavior fosters greater trust in the audit's assessments. Such insight can come from direct stakeholder engagement and through news stories, academic research, or other sources of information on AI's risks: what the UK's Information Commissioner's Office calls business intelligence [21]. Specific AI impacts for which systems need to be audited can also be discovered through complementary tools such as fundamental/human rights impact assessments, algorithmic impact assessments, and intersectional impact analyses conducted by policymakers in cognate domains

[112],[113],[114]. Criteria informed by these tools can allow auditors to evaluate whether the auditee has addressed known system harms.

Not every audit criterion is equal, however; some are better suited to a particular evaluation context than others. Selecting these criteria, then, requires a process that evaluates the relative usefulness of various evaluation metrics [56],[115]. These criteria should be able to accurately evaluate the alignment between what an AI system does and what it is expected to do. Setting criteria that establish a bright line between conformity and nonconformity allows auditors to make their decisions on firm, verifiable grounds.

d. **The auditor should clarify what constitutes an organization's conformity or compliance with a certain AI-related regulation, standard, or set of internal controls.**

Explanation: Evaluating conformity or compliance, which depends on the auditor's discretion, is often formalized as a binary proposition—a yes or no question. For example, the ISO 42001 standard simply asks whether bias tests have been conducted by the AI deployer [18]. It is more complicated to ask whether an AI system is subject to appropriate human oversight, as the EU's AI Act does [28]. The measure of compliance here may not be easily quantified [116]. In such cases, explaining what factors and information shaped the auditor's decision can demonstrate that it was well reasoned.

e. **When auditing an AI system's fairness, the criteria should include fairness metrics that evaluate parity in treatment and/or impact, especially for individuals from historically marginalized groups. The selected criteria should be appropriate for the sociocultural context of the audit. Bias mitigation techniques should also have been applied to the system data and constituent models.**

Explanation: Fairness metrics are used to assess whether the outputs of an AI system are unfairly biased against certain people [77]. Audit criteria that use these metrics often establish the threshold for what counts as a fair outcome at a decision-making juncture. It is difficult, however, to define a universal measure of fairness. Research across multiple

disciplines has produced several fairness metrics, each with its own specific, contextually valuable applications. These metrics are embedded with certain cultural, societal, and moral assumptions that shape how fairness is constructed and defined. As such, they may not account for varying global dynamics of power and fairness [31],[81]. Useful fairness criteria, then, are defined in relation to the group for which fairness is being measured: they pay attention to who faces the brunt of unfairness and what fairness means to them. It is also important for measures of system fairness to consider the impact of dataset quality and bias, which is a consequence of how data is constructed and how datasets are produced [118],[119]. Selecting fairness criteria that account for these contextual differences makes for a more effective audit. It is therefore vital for bias mitigation techniques to also take into account what counts as fairness in an AI system's sociocultural context.

f. **When auditing an AI system's transparency, the criteria should include whether the AI developer or deployer has clearly disclosed the following to those who use or are otherwise impacted by the system: documentation of data and model components and associated software and hardware resources; documentation of data collection, use, storage, and processing processes and associated labor resources; documentation of model features; and the expected role, behavior, and results of the system.**

Explanation: Transparency criteria can be used to assess what stakeholders can learn about the constituent parts of an AI system, including and beyond a system's technical features [78],[120]. Taking cues from the social transparency paradigm, such criteria can be used to assess whether the AI developer or deployer has clearly disclosed how the system is expected to work [121]. They are especially relevant when assessing how an AI system interacts with social structures and processes [122]. Transparency disclosures can happen through direct notifications, system artifacts, or accessible labels in language that nontechnical stakeholders can clearly understand [32],[33].

g. **When auditing an AI system's interpretability and explainability, the criteria should include whether the system's internal features and outcomes are explainable to stakeholders. It is also important that the quality of the explanations provided about the AI system (in such terms as clarity, comprehensiveness, and understandability) is acceptable.**

Explanation: An AI system's interpretability and explainability demonstrate that the system's algorithmic mechanics are inherently explainable and, where they are not, that the explanations provided about the system are clearly understood by stakeholders [40]. The system's post hoc explainability, which describes whether stakeholders can easily understand why a system produces the outputs that it does and what those outputs mean, is of particular importance [123]. Such explainability is a nonbinary characteristic that can be understood through the explanation's contents, its presentation format, and how it is presented to stakeholders—that is, through how a system is explained and how it is understood [124],[125].

h. **When auditing an AI system's robustness and safety, the criteria should consider whether the system developer has subjected the model(s) to adversarial training; appropriate cross-validation methods; minimization of concept drift and of overfitting to training data. These criteria should also account for whether the system developer has an appropriate emergency response and fallback plan in case of an attack; whether the system is regularly updated in response to changed data and operating conditions; and whether human-in-the-loop oversight has been incorporated into the system.**

Explanation: Audit criteria that denote an AI system's robustness and safety describe the system's ability to continue operating safely under adverse conditions [40]. Adversarial training and cross-validation are necessary defensive measures that can account for malicious inputs before deployment and help systems maintain accuracy in unstable environments. Mitigating concept drift and overfitting also helps maintain performance when

the system is transferred to an operating environment [126]. Oversight mechanisms are also crucial for ensuring that a system can be restored to normal functionality in case of a failure or breakdown. Robustness and safety criteria can therefore be used to measure whether developers and deployers have minimized the likelihood that an AI system can be adapted to malicious ends or engender outcomes harmful to public safety and security [31].

i. **When auditing system outputs for their accuracy, the criteria should specify the benchmarks against which accuracy is assessed and why they are relevant.**

Explanation: Accuracy benchmarks measure how closely an AI system's outputs align with "ground truths" represented in the benchmark dataset. The Pilot Parliaments Benchmark, for example, was developed to audit various facial recognition technologies (FRTs) for how accurately the system would recognize individuals from different demographic groups [70]. Audit criteria that rely on such benchmarks, then, can be used to evaluate whether system outputs align with the conditions stakeholders care about, such as truthfulness in AI-generated content.

j. **When auditing an AI system's training, test, and validation data, the criteria should assess compliance with data privacy and protection laws and regulations. Data collection, use, storage, processing, and sharing should also adhere to individual and collective data rights (such as the right to deletion) and to community-specific norms such as Indigenous data sovereignty. At the technical level, the collected data should have been subject to cleaning and debiasing. Easily understandable data documentation and provenance records should have been provided. In terms of data representativeness, the datasets should adequately represent local demographic groups and languages, especially if the datasets include synthetic data. The auditee should also demonstrate that they have obtained appropriate consent for data collection, use, and storage from stakeholder communities.**

k. **The auditor should engage with the communities from whom data is collected when deciding audit criteria.**

Explanation for j.–k.: Existing regulations, such as the EU's GDPR, offer a useful set of minimum protections for data collection, use, and storage that AI developers and deployers must adhere to. Privacy assessments conducted as part of an audit also demonstrate how rigorously AI developers and deployers address data privacy protection. The Office of the Privacy Commissioner of Canada (OPCC) has, for example, detailed what must be included in a privacy assessment through a set of ten "fair privacy principles". Such assessments are required by Canada's (2009) Personal Information Protection and Electronic Documents Act, the federal privacy law for the private sector [127]. Regulators such as the UK's Information Commissioner's Office and the U.S. Government Accountability Office include existing data-related regulation in their audit criteria [20],[21]. However, AI systems may also use data in ways that are inadequately addressed by existing regulations.

In the absence of appropriate regulatory frameworks, especially for communities with their own norms around data, audit criteria cannot be limited to compliance. Data, much like other forms of recording specific details of people's lives, is informed by social arrangements and cultural norms [128]. Over- or under-representation of certain cultures, groups, and realities in the training and testing data can affect an AI system's fairness and accuracy [129],[130]. Since the data used in an AI system directly affects its outcomes, developers must ensure that various stakeholders decide what data about them is collected and how it is collected, used, stored, shared, and deleted.

Auditees' adherence to individual data rights ensures that the individual can control how their personal data is processed [131]. Ensuring collective data sovereignty allows communities to collectively assert power over data about them [37],[132]. These norms also inform what data should not be collected. Data sovereignty is especially important for protecting the rights of Indigenous communities, who have historically been subject to

violent data extraction and erasure, to have control over their relationship to data [133],[134].

l. **When auditing an AI developer's or deployer's commitment to fair labor practices, the criteria should consider labor laws, regulations, and workers' rights. Additional criteria can account for whether data and system maintenance workers are paid at least the local minimum wage, guaranteed secure employment through fair contracts, and have adequate protections from physical and mental risks. AI deployers should also obtain worker consent through opt-in mechanisms for using, collecting, storing, sharing, and deleting data about them. Adherence to workers' rights also requires AI developers to institute mechanisms for workers to contest automated decision-making and task automation. Where feasible, AI deployers should also ensure that workers are provided compensation for displaced work and have access to reskilling initiatives.**

m. **Workers should be engaged in developing criteria related to their role in developing, deploying, and maintaining AI systems, including through worker collectives such as unions and through civil society organizations that take up workers' interests.**

Explanation for l.–m.: AI systems affect workers throughout the AI supply chain and life cycle, from data labor to infrastructural maintenance, and might lead to loss of job opportunities [135],[136]. Which workers are affected by AI systems and how they are affected also depends on their jurisdictional, geographical, and social location. Many of AI's impacts are felt most acutely by workers in majority world countries, where economic opportunities are already precarious and limited [137]. Audit criteria must therefore account for AI's impacts on labor throughout its supply chain and life cycle. Existing labor regulations offer minimum criteria for assessing whether workers have adequate protection against the adverse impacts of AI systems. Additionally, principles relating to labor and AI—such as ensuring fair production networks—have been developed by nonprofit organizations

specifically concerned with labor, such as FairWork [138]. Audit criteria that formalize these principles can help protect workers.

Given the difficulty auditors face when trying to identify the impacts on workers within the AI supply chain and life cycle—especially given how auditees can obscure areas of concern therein—workers themselves are best placed to direct auditors' attention to problematic aspects of the system. Workers and civil society organizations can also collaborate with auditors to systematize and formalize the concerns of a particular labor sector as audit criteria.

n. **When auditing an AI system's environmental sustainability, the criteria should take into account environmental laws and regulations and community needs. Related criteria should also take the following into account: minimization of the system's energy and water consumption, including through the choice of model and data-processing techniques; minimization of greenhouse gas emissions; minimization of carbon footprint and CO_2 inefficiency; environmental certifications for computational hardware and data centers; hardware reuse and recycling; and safe e-waste disposal.**

o. **Environmentalists, scientific experts, and communities impacted by AI should be engaged in developing audit criteria related to a system's environmental sustainability.**

Explanation: AI development and usage have amplified many of the impacts of the existing technical infrastructure, such as driving significant spikes in energy and water usage while training LLMs and at data centers used by AI companies. Drastic increases in chip production, essential to AI's growth, also fuel energy consumption and environmental pollution through mining and production processes [139],[140]. AI systems are also used by fossil fuel and mining companies to speculate on the location of untapped reserves, which can drive environmentally damaging resource extraction [4]. Existing environmental law offers a

minimum set of audit criteria for AI's environmental linkages, motivating the assessment of concerns such as resource sustainability.

While existing research has identified certain key indicators for AI's environmental impacts, audits can be used to assess whether AI developers and deployers have taken steps to minimize those impacts [141],[142]. The globally cumulative—rather than localized—environmental impacts of AI systems require auditors to incorporate input from the full range of stakeholders affected by a system [143]. Audit criteria informed by scientific expertise can help systematize disparate knowledge of communities and civil society, accurately pointing to areas of assessment that local regulation might miss [52],[101].

p. **When auditing the social impacts of an AI system for which well-established audit criteria do not exist, such criteria should be developed in accordance with stakeholders' expectations of impact mitigation.**

q. **When reusing criteria developed for one AI audit in another audit, the auditor should account for differences in the kind of system being evaluated and its local impacts.**

Explanation for p.–q.: What defines social harms can vary depending on socioeconomic status and cultural conditions, as well as the historical and contemporary power imbalances across different countries [144],[145]. Furthermore, an AI developer's or deployer's perspective on their systems cannot alone adequately account for their real-world operations, especially when the developer or deployer is located in wealthier countries such as the United States or the United Kingdom. Impacted individuals and communities, though, are usually only able to identify a system's post-deployment impacts. The perceptions and experiences of AI impacts also differ significantly among local stakeholders.

Consequently, auditors will need to carefully translate AI's impacts into context-sensitive impact metrics or conformity checks based on what can be measured within a particular context [90]. Audit criteria developed at the interface between sets of stakeholders can clarify how an AI developer's impact mitigation strategies can be evaluated. Local multi-stakeholder

audit design processes enable the development of criteria that translate disparate systemic impacts into the particular concerns of local stakeholders. This can be especially useful in jurisdictions where AI governance efforts are not in place or are inadequate regarding AI's impacts in that area.

6.5. Evidence

Evaluating a system against the audit criteria detailed in 6.4. requires comprehensive evidence. Our recommendations for ensuring auditors can gather the evidence they need are as follows:

a. **AI developers or deployers should provide auditors with the evidence they request, especially model cards, datasheets, audit trails and/or logs, and data documentation such as data provenance records.**

Explanation: AI developers and deployers with an internal audit function—a vital assurance mechanism—gather information on potential system failures or malfunctions reported by users. Such incident records can serve as a valuable source of audit evidence and can demonstrate the effectiveness and reliability of internal auditing processes [146]. Reliably providing comprehensive system and data documentation to auditors—irrespective of their relationship to the auditee—is essential for a rigorous audit [147]. In social audits, for example, government entities are required to provide community auditors with access to government records and sites built with or supported by public funds [92]. Voluntary disclosure regimes that provide comprehensive system artifacts, such as the system cards shared by OpenAI and Meta, also furnish auditors with evidence they can use to inform their audit [10],[148],[149]. Proper access to evidence allows auditors to evaluate whether the auditee has accounted for potential risks and impacts across the AI supply chain and life cycle. It also allows independent auditors to accurately determine what data might be missing from AI developers' historical datasets and how to complement them [60].

b. **Auditors and auditees should agree on the data sharing mechanisms that must be in place between them while they are sharing proprietary, sensitive, or confidential data. These mechanisms should also adhere to trust, identity, privacy, and security considerations. Where possible, trusted intermediaries such as data observatories, data commons, and data cooperatives should facilitate data sharing.**

Explanation: Secure data sharing platforms and protocols, such as those employed by cybersecurity professionals, can help mitigate auditee concerns around leaks while furnishing auditors with the data they need to conduct an effective audit [150],[151]. These technical mediums must be developed by AI developers and deployers in collaboration with auditors and other stakeholders such as users and impacted communities. Data sharing infrastructure can also facilitate auditors' data access by paying due attention to safety and privacy concerns. Only the minimum data required for an effective audit (to which security measures such as anonymization are applied) needs to be disclosed to the auditor. Considering whether data is shared through "secure and trusted systems for delivering privacy and digital identity management" is also vital when handling sensitive information, and principles articulated in domains such as medical research are relied on [152, p. 3], [153].

Enabling data access often presents significant logistical challenges. Improper handling of data received from AI developers or deployers can expose proprietary or confidential information to adversarial actors, jeopardizing the auditee's systems. Additionally, leaks of sensitive or personal data may lead to privacy violations. Data access becomes more complex when parties in the AI life cycle or supply chain are located across different jurisdictions [67]. This is especially the case when an AI deployer located in the majority world, such as a local government entity, uses systems developed in the minority world. Auditing such systems can impose undue burdens on the AI deployer.

Trusted intermediaries, such as international data observatories, can help address some of these challenges. They can facilitate the secure collection, use, and storage of audit evidence

[154], [155]. They can also help manage cross-border data flows according to regulations. Consistent data collection, curation, and storage practices at institutional sites grant auditors systematic access to datasets used by AI developers, even as AI systems change. Maintaining detailed provenance records for datasets can help ensure transparency regarding data context, transformations, and manipulations. Standardizing data collection can help facilitate data reuse for multiple audits while ensuring compliance with legal standards for sharing data across jurisdictions and stakeholders.

c. **Where relevant, audit evidence should be systematically collected from end-users and communities impacted by an AI system.**

Explanation: Professional auditors—internal and external—are often unable to directly monitor an AI system's development processes or collect data about how a system operates in real-world conditions. Auditors have historically sought to resolve these issues through site visits and interviews with stakeholders that have direct insight into these processes. Such techniques for evidence collection are, however, of limited use when auditing AI systems whose impacts extend beyond the auditee [143],[144]. End-users of AI systems and communities impacted by them are often able to directly collect accurate and contextually sensitive—if unsystematized—evidence of AI's operations [65]. (Data) workers' inquiries, for example, provide otherwise unavailable insights into how AI systems affect the workers involved in their development [156]. Community members affected by AI can use data collection instruments too, such as worksheets, to elaborate how these systems enter their daily lives [157].

Given the global and technologically mediated dispersion of AI users and affected communities, however, the evidence they collect (often anecdotally) needs to be systematized to be valuable in an audit. Expert auditors typically perform this translation work, often with the help of automated tools [158],[159]. Such evidence is comparable to data and documentation provided by the AI developer or deployer.

d. **When auditing for cumulative impacts, audit evidence should be collected over an extended period and from multiple sites.**

Explanation: Longitudinal evidence collection is vital when the audited AI system's behavior changes frequently in response to input data or real-world conditions [60]. This is often the case with large language models and other generative systems that are used across the world. These systems can also change based on the geographical region and application domain in which they are used, making multi-sited audits valuable [160]. Conducting audits across time and place also produces evidence of disparate global impacts. For example, algorithmic fare-setting on ride-sharing platforms may initially appear fair but can lead to income inequities over time and across different driver locations [161],[162].

6.6. Methodology

Recommendations for an audit methodology that can effectively assess the audit evidence against established criteria are as follows:

a. **The auditor should conduct performance and scenario tests to evaluate whether an AI system meets established criteria for fairness, interpretability and explainability, robustness and safety, and accuracy. If direct testing is not feasible, the auditor should document the alternative methods used to collect, verify, and validate results, explaining why these methods were suitable substitutes.**

b. **The auditor should conduct performance and scenario tests appropriate to the system's model(s), training, testing, and validation data, its interaction domains, and its potential harms.**

c. **The auditor should assess the system's ability to deliver consistent outcomes within an application domain, accounting for edge cases and any violations of established guardrails.**

Explanation for a.–c.: Effective audits cannot rely on tests conducted by the AI developer themselves at the development or pre-deployment stage. These tests may not account for how users interact with AI systems under real-world operating conditions [70]. Furthermore, the results of performance tests cannot be validated through comparison with system documentation alone, since it is difficult to definitively attribute system outputs to certain model features or training data characteristics [163]. Such difficulties also arise in the case of "fine-tuned" models, where the resources and methods used to adapt a foundation model to a particular task can impact system performance [82].

Appropriate performance tests can account for the range of ways a system can be deployed and used, including when a user makes it perform in unanticipated ways. Such testing is especially important for general purpose AI systems meant to perform a variety of tasks, or when a system positioned as domain-agnostic performs better at some tasks than others [164]. For example, LLMs currently produce more accurate outputs in everyday communication than in code generation [165]. Auditors can also use toolkits such as Aequitas and the Adversarial Robustness Toolbox to subject a system to the test most appropriate to the audit [56],[166]. Scenario testing, such as through simulations, helps test an AI system's potential uses and behaviors in real-world conditions [161],[167],[168]. Audits that assess the risk of disparate treatment, for example, rely on counterfactual fairness tests, allowing the auditor to compare how an individual or group would have been treated under various circumstances [169]. Performance tests need not cover every use-case of a system; instead, the conducted tests should effectively account for the use-cases covered by the audit object(s) and criteria.

d. **The auditor should employ documentary analysis and testimonial gathering to evaluate the AI system's transparency. These methods are also useful for assessing compliance with data privacy and protection laws and the quality of data documentation and provenance records. They can also be used to evaluate AI developers' and deployers' commitment to fair labor practices, respect for human rights, and environmental sustainability.**

e. **Stakeholder interviews and site visits should be facilitated by the AI developer or deployer.**

Explanation for d.–e.: Audits have traditionally relied on documentary analyses and testimonial gathering since an auditee's records and their staff's insights often hold important information about an organization's operations. Documentation is also vital for describing system features and an organization's use and governance of a system. These methods enable auditors to assess parts of the AI supply chain that auditors cannot directly access themselves, such as a system's component suppliers. They can also help auditors review decisions made earlier in the AI life cycle, like design choices [18]. Furthermore, the widely accepted use of such methods can shift the evidentiary burden to AI developers and deployers, ensuring that auditors can produce comprehensive assessments of an AI system. At the same time, auditees' documentation of an AI system's development processes—such as self-attestation of fair labor practices—cannot be deemed reliable if auditors are unable to observe the same conditions in their investigations [90].

f. **Stakeholder engagement should, where possible, employ participatory methods to audit the AI system's fairness, interpretability, and explainability. Impacted stakeholders can provide insight into an AI developer's or deployer's commitment to cultural norms, Indigenous data sovereignty, fair labor practices, and environmental sustainability.**

g. **The auditor should demonstrate the value and suitability of their methodology when using bespoke or nonstandard methods and tools; collaborating with community and civil society stakeholders; or auditing systemic impacts of an AI system for which standard methods and tools have yet to be established.**

Explanation for f.–g.: Participatory or user-driven methods are particularly useful for assessing whether the ways people encounter AI systems in their everyday lives, including through indirect impacts such as labor automation, adhere to human rights, sociocultural

values, and regulatory requirements. Many sociotechnical and end-user audits rely on systematizing user reports of problematic system behavior [158],[170]. Other stakeholder-driven methods, such as data workers' inquiries, can also help auditors incorporate evidence from impacted communities beyond the application domain [156]. By centering those stakeholders who have a direct or indirect relation to a specific AI system, such methods also demonstrate a sensitivity to context that is extremely valuable when auditing a system's disparate global impacts. Stakeholder-engaged methods, then, bring into the audit people whose perspective might otherwise be lost. Such methods therefore enable auditors to produce more comprehensive assessments of a system than model- or auditee-centric methods [157].

h. **The auditor should conduct repeated or longitudinal audits to assess an AI system's accuracy, fairness, and robustness and the developer's or deployer's mitigation of cumulative or systemic impacts.**

Explanation: Repeating data collection processes at set intervals or applying continuous auditing processes can provide insight into constantly evolving AI systems [65]. The efficacy of such methods can be bolstered by tying them to practices such as system version control that allow developers or deployers to track changes in an AI system's behavior and functionality [40].

i. **The auditor should satisfy ethical obligations relevant to their methodology, such as having approval from an institutional or ethical review board (IRB/ERB).**

Explanation: Auditors may employ nonstandard procedures when addressing practical and ethical limitations of standard methodologies. End-user audits and participatory methods, for example, are useful when they hew closer to the concerns of the public and respect their sociocultural norms and values [158],[170]. Auditors may also employ alternative methodologies when conventional ones, such as desk audits, cannot access primary or secondary evidence or key stakeholders. For example, data annotation work is often conducted by a geographically dispersed workforce, making site visits unfeasible [2].

If stakeholders—the public, regulators, or AI developers—are to trust such an audit, though, its methodology must fulfill a fundamental requirement: it should be able to demonstrate a verifiable causal relationship between the audit evidence and the opinion issued therein [145]. Auditors who have carefully accounted for stakeholders concerns prior to conducting the audit are also trusted more [106].

j. **The auditor should institute mechanisms to verify stakeholder satisfaction with the selected audit features.**

Explanation: The role of stakeholders should not be limited to helping define the audit scope, criteria, or evidence collection [171]. Sustaining stakeholder engagement after the audit has been conducted, such as by properly communicating the findings, is essential to meeting auditing's social accountability goals. Auditors might also revise their opinion if stakeholders identify deficiencies in the audit. Measuring stakeholder satisfaction with the audit process can help show how well societal concerns are addressed throughout the audit. Demonstrating that stakeholder input makes a material difference to the audit's design and process enhances its trustworthiness [172].

6.7. Post-Audit Activities

Recommendations for conducting post-audit activities, including report development and submission are as follows:

a. **The audit report should clearly state the audit scope. It should clarify the role of the auditor and their relationship with the audited AI developer or deployer. It should specify which stakeholders were engaged during the audit and why. It should state what audit criteria were used. A description of the audit methodology should describe what evidence was collected and how, specifying the performance tests, analyzed documents, and stakeholder engagement methods. The results of these assessments should be clearly stated, along with any improvements made based on past audit findings and whether any new areas for improvement were identified.**

b. **When contracted by an AI developer or deployer to conduct a conformity assessment, the auditor should clearly state in the report how roles and responsibilities associated with an AI system's governance are distributed among the auditee's staff. In addition, the audit report should clearly state a point of contact for the AI developer or deployer in case stakeholders want to contest particular aspects of an AI system or request further information pursuant to the audit**

Explanation for a.–b.: A comprehensive final audit report addresses how the audit was designed and conducted, including descriptions of which of the considerations from this framework were applied [173]. Such a report helps assure stakeholders, including the auditee and the public, that an AI system aligns with their expectations of it. Stakeholders can also use the report to identify where AI developers or deployers need to make improvements to their systems [86]. Consequently, a well-justified positive audit opinion serves not only as a validation of the system itself but also as a means to reinforce the trustworthiness of the audited AI developer or deployer. The details of the report also allow stakeholders to properly assess the rigor and efficacy of the audit itself [18]. Finally, the report could also help stakeholders identify possible deficiencies in the audit that could be assessed as part of follow-up audits.

Recording which members of the organization are responsible for various aspects of an AI system also clarifies who outside the organization they are accountable to. For example, the Chief Legal Officer of an AI developer is answerable to external regulators for what processes and procedures are in place to ensure compliance.

c. **The audit report should first be submitted to the stakeholders with whom the auditor has accountability relationships.**

d. **Once stakeholders have been granted a reasonable time to review the audit report and address its most urgent findings, it should be made publicly available**

through widely accessible news media or searchable public repositories, including open access archives such as arXiv and OSF.

Explanation for c.–d.: Ensuring that the concerns raised during an AI audit are addressed requires those responsible for the corresponding aspects of an AI system to be properly informed. Audit committees within an auditee's organization typically manage the receipt of internal and third-party audit reports and the communication of relevant information to organizational stakeholders. With independent audits, however, proper mechanisms are required to ensure that the auditee has clear and secure information on potential flaws or risks of their system. Such information must not be disclosed to adversarial actors who might exploit it but should be channeled to those employees of an AI developer or deployer who can address the problem area(s) quickly and effectively. Algorithmic bug bounties, coordinated flaw disclosure, and other incident reporting mechanisms can make audit results systematically available to the relevant stakeholders [174],[175].

Making audit reports globally accessible, such as through international audit databases, can also help make audits conducted in one region useful to stakeholders elsewhere [23]. This can minimize the regulatory burden on auditees. A widely accessible audit report that demonstrates the environmental sustainability of data centers used by multiple AI developers, for example, minimizes the need for redundant audits of those data centers for each of those developers. Enabling such transparency, however, is the minimum condition for AI audits to usefully fulfill an accountability function**.**

e. **If an audit finds that the auditee is responsible for negative impacts or harms engendered by their AI system, the auditee should take adequate steps to improve their system and/or its governance, given adequate safe harbor and a reasonable grace period.**

f. **The audit report should identify areas of concern to be addressed in follow-up audit(s).**

Explanation for e.–f.: AI developers' and deployers' public accountability need not be excessively—or even primarily—punitive. Audit reports are first and foremost produced so that auditees can identify and correct deficiencies, but audits might discover system harms that invite litigation or liability. Providing auditees with adequate time to address those issues ensures harm mitigation without resorting to expensive legal procedures or causing reputational losses for the AI developer or deployer. Nonpunitive forms of accountability also encourage AI developers and deployers to be transparent about aspects of their systems that might be noncompliant or harmful rather than attempting to obfuscate them from auditors [176]. Actionable auditors' opinions and proactive, timely responses by auditees, then, can help maintain trust between these parties and public trust in the effectiveness of the audit ecosystem. The ecosystem can also be supported by the future development of independent accountability mechanisms, such as through regulatory enforcement or an oversight board that is independent of the AI market but institutionally supported.

7. CONCLUSION

The global AI auditing framework proposed here provides a practical approach to **effective audits of AI systems with disparate global impacts**. These recommendations translate consensus on AI auditing design and practices into actionable steps, addressing gaps in current audit practices. The framework considers contextual and stakeholder-driven factors, including audit scope, the auditor's role (such as independence and qualifications), criteria and evidence, methodology, and post-audit activities. It also examines how, on one hand, local impacts of AI systems contribute to systemic risks and harms and how, on the other hand, global impacts manifest in local contexts.

We note here that despite progress in technological development, **asymmetries persist between stakeholders in majority world countries and those in other parts of the world** that shape how AI audits can be conducted**.** Most existing audit regimes prioritize compliance with regulatory and sociocultural norms in English-speaking minority world countries [16], meaning that these norms heavily influence current definitions of audit scope and methodology. While existing AI audit criteria, tools, and methods are beginning to address social and environmental concerns relevant to majority world communities, critical concerns such as Indigenous data sovereignty and labor precarity remain underrepresented in audit discourse. Furthermore, the uneven global distribution of accountability mechanisms, data sharing agreements, and cross-jurisdictional policies on data privacy, protection, and sovereignty complicates the execution of effective global AI audits.

This framework therefore prioritizes the perspectives of communities that are most vulnerable to AI harms rather than imposing generic concerns onto local contexts [72],[177]. Simultaneously, it allows auditors and auditees in minority world countries to tailor audit requirements to their local contexts, thereby avoiding the burdens of a rigid set of universal standards.

AI audits aligned with our recommendations can help **develop an audit ecosystem that fosters public trust** by providing the grounds for public assurance. Rather than privilege a single kind of audit, our framework enables AI systems to be assessed from multiple perspectives while ensuring procedural regularity and quality. It also provides mechanisms for stakeholders to assess and contest the outcomes of AI systems using information gathered during an audit, empowering them to hold AI developers and deployers accountable.

Our recommendations highlight the limitations of existing auditing research and practices, thereby demonstrating the need for **institutional capacity-building and continued academic research into AI audit design and practices.** For instance, our recommendations for proper data sharing mechanisms depend on the continuous improvement of related scientific fields and the development of trusted international intermediaries that can facilitate auditors' access to data.

The growing risks AI systems pose to social, economic, cultural, political, and environmental systems and processes require an expanded set of audit practices, many of which are still under development. While current audit criteria address known risks and impacts, they fall short when it comes to mitigating harms and ensuring public safety. Furthermore, existing audit methodologies, including evidence collection and evaluation mechanisms, largely borrow from non-AI domains, limiting their effectiveness in addressing AI-specific challenges. Auditors employing such innovative practices also need support from organizations that can provide the financial and informational resources needed to achieve the goals of the audit.

To improve audits, we need more creative and stakeholder-engaged approaches to flexibly address the globally dispersed, still-evolving nature of AI systems. This includes tackling the challenges of measuring difficult outcomes, such as the proliferation of harmful content and the reification of certain sociocultural norms. Continued academic research is crucial to overcoming these challenges and enhancing the effectiveness of AI audits.

REFERENCES

[1] W. D. Heaven, "Google's medical AI was super accurate in a lab. Real life was a different story.," *MIT Technology Review*, Apr. 27, 2020. https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/

[2] J. Dzieza, "Inside the AI Factory," *The Verge*, Jun. 20, 2023. https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots

[3] Edward Ongweso Jr, "The Lockout: Why Uber Drivers in NYC Are Sleeping in Their Cars," *VICE*, Mar. 19, 2020. https://www.vice.com/en/article/the-lockout-why-uber-drivers-in-nyc-are-sleeping-in-their-cars/.

[4] M. Hogan, "The Fumes of AI," *Critical AI*, vol. 2, no. 1, Apr. 2024, doi: https://doi.org/10.1215/2834703x-11205231.

[5] C. Qu and E. Kim, "Reviewing the Roles of AI-Integrated Technologies in Sustainable Supply Chain Management: Research Propositions and a Framework for Future Directions," *Sustainability*, vol. 16, no. 14, p. 6186, Jan. 2024, doi: 10.3390/su16146186.

[6] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: a three-layered approach," *AI Ethics*, May 2023, doi: 10.1007/s43681-023-00289-2.

[7] B. Faveri and G. Marchant, "International Artificial Intelligence (AI) and AI-Related Standards Landscape: ISO, IEEE, ETSI, and ITU," 2024, Available: https://doi.org/10.7910/DVN/LBIXU6

[8] G. Auld, A. Casovan, A. Clarke, and B. Faveri, "Governing AI through ethical standards: learning from the experiences of other private governance initiatives," *Journal of European Public Policy*, vol. 29, no. 11, pp. 1822–1844, Nov. 2022, doi: 10.1080/13501763.2022.2099449.

[9] UNESCO, "UNESCO and the Thomson Reuters Foundation partner to launch the AI Governance Disclosure Initiative," 2024. https://www.unesco.org/en/articles/unesco-and-thomson-reuters-foundation-partner-launch-ai-governance-disclosure-initiative (accessed Nov. 21, 2024).

[10] Forum on Information and Democracy, "A Voluntary Certification Mechanism for Public Interest AI: Exploring the Design and Specifications of Trustworthy Global Institutions to Govern AI," Forum on Information and Democracy, Sep. 2024. Available: https://informationdemocracy.org/wp-content/uploads/2024/09/FID-Public-Interest-AI-Sept-2024.pdf

[11] E. Roth, "AI landlord screening tool will stop scoring low-income tenants after discrimination suit," *The Verge*, Nov. 20, 2024. https://www.theverge.com/2024/11/20/24297692/ai-landlord-tool-saferent-low-income-tenants-discrimination-settlement.

[12] International Panel on the Information Environment, 2024. *Global Approaches to Auditing Artificial Intelligence: A Literature Review*. SR2024.1. Zurich, Switzerland: IPIE.

[13] A. L. Watkins, W. Hillison, and S. E. Morecroft, "Audit Quality: A Synthesis of Theory and Empirical Evidence," *Journal of Accounting Literature*, vol. 23, pp. 153–193, 2004, Available: https://www.proquest.com/openview/848c22290cbd79f08e07f1a6a491974d/1?pq-origsite=gscholar%26cbl=31366

[14] B. Faveri and G. Auld, "Informing Possible Futures for the use of Third-Party Audits in AI Regulations," Carleton University, Regulatory Governance Initiative, Dec. 2023. [Online]. Available: http://doi.org/10.22215/sppa-rgi-nov2023

[15] A. Urman, M. Makhortykh, and A. Hannak, "Mapping the Field of Algorithm Auditing: A Systematic Literature Review Identifying Research Trends, Linguistic and Geographical Disparities," 2024, *arXiv*. doi: 10.48550/ARXIV.2401.11194.

[16] M. Sloane, E. Moss, and R. Chowdhury, "A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability," *Patterns*, vol. 3, no. 2, Art. no. 2, Feb. 2022, doi: 10.1016/j.patter.2021.100425.

[17] M. Sloane, "To make AI fair, here's what we must learn to do," *Nature*, vol. 605, no. 9, May 2022, doi: https://doi.org/10.1038/d41586-022-01202-3.

[18] Information technology - artificial intelligence - management system, ISO Standard No. 42001:2023, 2023. Available: https://www.iso.org/standard/44546.html.

[19] New York City Consumer and Worker Protection, Local Law 144. 2021. [Online]. Available: https://rules.cityofnewyork.us/wp-content/uploads/2023/04/DCWP-NOA-for-Use-of-Automated-Employment-Decisionmaking-Tools-2.pdf

[20] U.S. Government Accountability Office, Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. GAO-21-519SP, 2021. [Online]. Available: https://www.gao.gov/assets/gao-21-519sp.pdf

[21] Information Commissioner's Office, A guide to ICO audit: Artificial Intelligence (AI) audits. 2022. [Online]. Available: https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf

[22] "GPAI and OECD unite to advance coordinated international efforts for trustworthy AI," *OECD*, Jul. 03, 2024. https://www.oecd.org/en/about/news/speech-statements/2024/07/GPAI-and-OECD-unite-to-advance-coordinated-international-efforts-for-trustworthy-AI.html

[23] G. Falco et al., "Governing AI safety through independent audits," Nat Mach Intell, vol. 3, no. 7, pp. 566–571, Jul. 2021, doi: 10.1038/s42256-021-00370-7.

[24] J. Mökander, "Auditing of AI: Legal, Ethical and Technical Approaches," *DISO*, vol. 2, no. 3, p. 49, Nov. 2023, doi: 10.1007/s44206-023-00074-y.

[25] Aparna Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi, "The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations," *2022 ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2022, doi: https://doi.org/10.1145/3531146.3533179.

[26] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," *ACM Computing Surveys*, Aug. 2023, doi: https://doi.org/10.1145/3616865.

[27] I. D. Raji et al., "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: ACM, Jan. 2020, pp. 33–44. doi: 10.1145/3351095.3372873.

[28] L. Floridi, M. Holweg, M. Taddeo, J. Amaya Silva, J. Mökander, and Y. Wen, "capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act," SSRN Electronic Journal, 2022, doi: https://doi.org/10.2139/ssrn.4064091.

[29] K. Lam, B. Lange, Borhane Blili-Hamelin, J. Davidovic, S. Brown, and A. Hasan, "A Framework for Assurance Audits of Algorithmic Systems," Jun. 2024, doi: https://doi.org/10.1145/3630106.3658957.

[30] J. Mökander, M. Axente, F. Casolari, and L. Floridi, "Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation," Minds & Machines, vol. 32, no. 2, pp. 241–268, Jun. 2022, doi: 10.1007/s11023-021-09577-4.

[31] L. Oala et al., "ML4H Auditing: From Paper to Practice," in Proceedings of the Machine Learning for Health NeurIPS Workshop, PMLR, Nov. 2020, pp. 280–317. Accessed: Jan. 18, 2024. [Online]. Available: https://proceedings.mlr.press/v136/oala20a.html

[32] S. Costanza-Chock, I. D. Raji, and J. Buolamwini, "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, in FAccT '22. New York, NY, USA: Association for Computing Machinery, Jun. 2022, pp. 1571–1583. doi: 10.1145/3531146.3533213.

[33] I. Ajunwa, "Automated Employment Discrimination," *Harvard Journal of Law & Technology*, vol. 34, no. 2, pp. 1–80, 2021, doi: https://doi.org/10.2139/ssrn.3437631.

[34] AI Now Institute, "Algorithmic Accountability: Moving Beyond Audits," *AI Now Institute*, Apr. 11, 2023. https://ainowinstitute.org/publication/algorithmic-accountability#weak-policy-response (accessed Nov. 21, 2024).

[35] R. Rodrigues, "Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities," *Journal of Responsible Technology*, vol. 4, no. 100005, pp. 1–12, Oct. 2020, doi: https://doi.org/10.1016/j.jrt.2020.100005.

[36] M. Latonero, "Governing artificial intelligence: Upholding human rights & dignity," Data & Society, 2018. Available: https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

[37] P. Hummel, M. Braun, M. Tretter, and P. Dabrock, "Data sovereignty: A review," *Big Data & Society*, vol. 8, no. 1, pp. 1–17, Jan. 2021, doi: https://doi.org/10.1177/2053951720982012.

[38] J. Wei, J. Zhao, and X. Hou, "Bilateral information sharing in two supply chains with complementary products," *Applied Mathematical Modelling*, vol. 72, pp. 28–49, Aug. 2019, doi: https://doi.org/10.1016/j.apm.2019.03.015.

[39] A. Belal, C. Spence, C. Carter, and J. Zhu, "The Big 4 in Bangladesh: caught between the global and the local," *Accounting, Auditing & Accountability Journal*, vol. 30, no. 1, pp. 145–163, Jan. 2017, doi: https://doi.org/10.1108/aaaj-10-2014-1840.

[40] A. Koshiyama et al., "Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms," Jan. 01, 2021, Rochester, NY: 3778998. doi: 10.2139/ssrn.3778998

[41] Public Company Accounting Oversight Board, "AS 2605: Consideration of the Internal Audit Function," *PCAOB*, 2023. https://pcaobus.org/oversight/standards/auditing-standards/details/as-2605-consideration-of-the-internal-audit-function_1528

[42] Institute of Internal Auditors, "1100 – Independence and Objectivity," 2024. https://www.theiia.org/en/content/guidance/recommended/implementation/1100-independence-and-objectivity/ (accessed Oct. 11, 2024).

[43] ForHumanity, "FORHUMANITY CERTIFIED AUDITORS (FHCA) CODE OF ETHICS," ForHumanity. Available: https://forhumanity.center/web/wp-content/uploads/2022/02/ForHumanity.center_FHCA_Code_of_Ethics.pdf

[44] P. W. Wolnizer, *Auditing as independent authentication*. 2006.

[45] M. Anderljung et al., "Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework," Nov. 15, 2023, arXiv: arXiv:2311.14711. [Online]. Available: http://arxiv.org/abs/2311.14711

[46] International Association of Algorithmic Auditors, "Become a member - International Algorithmic Auditors Association," *International Algorithmic Auditors Association*, Aug. 18, 2024. https://iaaa-algorithmicauditors.org/become-a-member (accessed Oct. 11, 2024).

[47] BABL AI, "AI and Algorithm Auditor Certification," *Babl.ai*, 2024. https://courses.babl.ai/p/ai-and-algorithm-auditor-certification?affcode=616760_ujptkhyg (accessed Oct. 11, 2024).

[48] J. R. Francis, "What do we know about audit quality?," *The British Accounting Review*, vol. 36, no. 4, pp. 345–368, Dec. 2004, doi: https://doi.org/10.1016/j.bar.2004.09.003.

[49] C. Wilson et al., "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: ACM, Mar. 2021, pp. 666–677. doi: 10.1145/3442188.3445928.

[50] BABL AI, "Summary of Bias Audit Results: Audit of Eightfold's Matching Model for New York City's Local Law 144," Apr. 2023. Available: https://eightfold.ai/wp-content/uploads/eightfold-summary-of-bias-audit-results.pdf

[51] E. P. Goodman and J. Tréhu, "AI Audit-Washing and Accountability" German Marshall Fund of the United States, Nov. 15, 2022. https://www.gmfus.org/news/ai-audit-washing-and-accountability

[52] A. Hilliard, A. Gulley, A. Koshiyama, and E. Kazim, "Bias audit laws: how effective are they at preventing bias in automated employment decision tools?," *International Review of Law Computers & Technology*, pp. 1–17, Sep. 2024, doi: https://doi.org/10.1080/13600869.2024.2403053.

[53] L. Stark, D. Greene, and A. L. Hoffmann, "Critical Perspectives on Governance Mechanisms for AI/ML Systems," in *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, J. Roberge and M. Castelle, Eds., Palgrave Macmillan Cham, 2021, pp. 257–280.

[54] R. Antle, "Auditor Independence," *Journal of Accounting Research*, vol. 22, no. 1, p. 1, 1984, doi: https://doi.org/10.2307/2490699.

[55] G. Filippi, S. Zannone, A. Hilliard, and A. Koshiyama, "Local Law 144: A Critical Analysis of Regression Metrics," Feb. 08, 2023, *arXiv*: arXiv:2302.04119. [Online]. Available: http://arxiv.org/abs/2302.04119

[56] P. Saleiro et al., "Aequitas: A Bias and Fairness Audit Toolkit," Apr. 29, 2019, *arXiv*:1811.05577. Available: http://arxiv.org/abs/1811.05577

[57] H. H. Chang, M. Bui, and C. McIlwain. "Targeted Ads and/as Racial Discrimination: Exploring Trends in New York City Ads for College Scholarships," 2023. *arXiv:*2109.15294.

[58] Charlton McIlwain. *Algorithmic Discrimination: A Framework and Approach to Auditing & Measuring the Impact of Race-Targeted Digital Advertising,* 2023*.* PolicyLink

[59] E. Purificato and E. W. De Luca, "What Are We Missing in Algorithmic Fairness? Discussing Open Challenges for Fairness Analysis in User Profiling with Graph Neural Networks," *Communications in Computer and Information Science*, pp. 169–175, 2023, doi: https://doi.org/10.1007/978-3-031-37249-0_14.

[60] M. Aidinoff et al., AI Auditing, unpublished

[61] V. K. Felkner, J. A. Thompson, and J. May, "GPT is Not an Annotator: The Necessity of Human Annotation in Fairness Benchmark Construction," *arXiv*, 2024, Available: https://arxiv.org/pdf/2405.15760

[62] L. Groves, J. Metcalf, A. Kennedy, B. Vecchione, and A. Strait, "Auditing Work: Exploring the New York City algorithmic bias audit regime," in The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro Brazil: ACM, Jun. 2024, pp. 1107–1120. doi: 10.1145/3630106.3658959.

[63] F. Gualdi and A. Cordella, "Theorizing the regulation of generative AI: lessons learned from Italy's ban on ChatGPT," in *Proceedings of the 57th Hawaii International Conference on System Sciences* , 2024, pp. 2023–2032. Available: https://hdl.handle.net/10125/106631

[64] E. Ferrara, "GenAI against humanity: nefarious applications of generative artificial intelligence and large language models," *Journal of Computational Social Science*, Feb. 2024, doi: https://doi.org/10.1007/s42001-024-00250-1.

[65] D. Metaxa et al., "Auditing Algorithms: Understanding Algorithmic Systems from the Outside In," FNT in Human–Computer Interaction, vol. 14, no. 4, pp. 272–344, 2021, doi: 10.1561/1100000083

[66] L. Wright et al., "Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability," in The 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro Brazil: ACM, Jun. 2024,pp. 1701–1713. doi: 10.1145/3630106.3658998

[67] J. Ford and J. Nolan, "Regulating transparency on human rights and modern slavery in corporate supply chains: the discrepancy between human rights due diligence and the social audit," *Australian Journal of Human Rights*, vol. 26, no. 1, pp. 27–45, Jan. 2020, doi: 10.1080/1323238X.2020.1761633.

[68] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: https://doi.org/10.1038/s42256-019-0088-2.

[69] L. Schmitt, "Mapping global AI governance: a nascent regime in a fragmented landscape," *AI and Ethics*, vol. 2, Aug. 2021, doi: https://doi.org/10.1007/s43681-021-00083-y.

[70] I. D. Raji, P. Xu, C. Honigsberg, and D. Ho, "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: ACM, Jul. 2022, pp. 557–571. doi: 10.1145/3514094.3534181.

[71] W. R. Knechel, "Do Auditing Standards Matter?," *Current Issues in Auditing*, vol. 7, no. 2, pp. A1–A16, Dec. 2013, doi: 10.2308/ciia-50499.

[72] J. Malcolm, "Criteria of meaningful stakeholder inclusion in internet governance," *Internet Policy Review*, vol. 4, no. 4, Dec. 2015, doi: 10.14763/2015.4.391.

[73] M.-T. Png, "At the Tensions of South and North: Critical Roles of Majority world Stakeholders in AI Governance," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul Republic of Korea: ACM, Jun. 2022, pp. 1434–1445. doi: 10.1145/3531146.3533200.

[74] E. Bondi, L. Xu, D. Acosta-Navas, and J. A. Killian, "Envisioning Communities: A Participatory Approach Towards AI for Social Good," *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, May 2021, doi: https://doi.org/10.1145/3461702.3462612.

[75] Y. Liu *et al.*, "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment," 2023. Available: https://arxiv.org/pdf/2308.05374

[76] G. Grill, "Constructing Certainty in Machine Learning: On the performativity of testing and its hold on the future", 2022. Available: osf.io/zekqv.

[77] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[78] R. Bommasani et al., "The Foundation Model Transparency Index," 2023, arXiv. doi: 10.48550/ARXIV.2310.12941

[79] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," Sci. Robot., vol. 4, no. 37, p. eaay7120, Dec. 2019, doi: 10.1126/scirobotics.aay7120.

[80] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat Mach Intell, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[81] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining Algorithmic Fairness in India and beyond," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 315–328, Mar. 2021, doi: https://doi.org/10.1145/3442188.3445896.

[82] R. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models," Jul. 12, 2022, *arXiv*: arXiv:2108.07258. Accessed: Jul. 24, 2024. [Online]. Available: http://arxiv.org/abs/2108.07258

[83] S. Longpre *et al.*, "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI," Nov. 04, 2023, *arXiv*: arXiv:2310.16787. Accessed: May 16, 2024. [Online]. Available: http://arxiv.org/abs/2310.16787

[84] C. Song and V. Shmatikov, "Auditing Data Provenance in Text-Generation Models," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2019. doi: https://doi.org/10.1145/3292500.3330885.

[85] Y. Huang, C.-Y. Huang, X. Li, and K. Li, "A Dataset Auditing Method for Collaboratively Trained Machine Learning Models," in *IEEE Transactions on Medical Imaging*, Institute of Electrical and Electronics Engineers, Nov. 2022, pp. 2081–2090. doi: https://doi.org/10.1109/tmi.2022.3220706.

[86] A. Birhane, R. Steed, V. Ojewale, B. Vecchione, and I. D. Raji, "AI auditing: The Broken Bus on the Road to AI Accountability," Jan. 25, 2024, *arXiv*: arXiv:2401.14462. Accessed: Feb. 29, 2024. [Online]. Available: http://arxiv.org/abs/2401.14462

[87] R. Shelby et al., "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction," in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Montreal QC Canada: ACM, Aug. 2023, pp. 723–741. doi: 10.1145/3600211.3604673.

[88] A. Delfanti and B. Frey, "Humanly Extended Automation or the Future of Work Seen through Amazon Patents," Science, Technology, & Human Values, vol. 46, no. 3, pp. 655–682, May 2021, doi: 10.1177/0162243920943665.

[89] S. Fox, S. Shorey, E. Y. Kang, D. A. Montiel, and E. C. Rodríguez, "Patchwork: The Hidden, Human Labor of AI Integration within Essential Work," in *Proceedings of the ACM on human-computer interaction*, Association for Computing Machinery, Apr. 2023, pp. 1–20. doi: https://doi.org/10.1145/3579514.

[90] G. Lebaron and J. Lister, "Benchmarking global supply chains: the power of the 'ethical audit' regime," *Rev. Int. Stud.*, vol. 41, no. 5, pp. 905–924, Dec. 2015, doi: 10.1017/S0260210515000388.

[91] S. McCracken, S. E. Salterio, and M. Gibbins, "Auditor–client management relationships and roles in negotiating financial reporting," *Accounting, Organizations and Society*, vol. 33, no. 4–5, pp. 362–383, May 2008, doi: https://doi.org/10.1016/j.aos.2007.09.002.

[92] Y. Aiyar and S. Samji, "Transparency and Accountability in NREGA: A Case Study of Andhra Pradesh," Accountability Initiative. Available: https://accountabilityindia.in/sites/default/files/working-paper/31_1244199489.pdf

[93] F. Afridi and V. Iversen, "Social Audits and MGNREGA Delivery: Lessons from Andhra Pradesh," *IZA Discussion Paper* No. 8095, 2014. Available at SSRN: https://ssrn.com/abstract=2424194 or http://dx.doi.org/10.2139/ssrn.2424194

[94] H. Chung, C. H. Sonu, Y. Zang, and J.-H. Choi, "Opinion Shopping to Avoid a Going Concern Audit Opinion and Subsequent Audit Quality," *AUDITING: A Journal of Practice & Theory*, vol. 38, no. 2, pp. 101–123, May 2019, doi: https://doi.org/10.2308/ajpt-52154.

[95] J. Stewart and N. Subramaniam, "Internal audit independence and objectivity: emerging research opportunities," *Managerial Auditing Journal*, vol. 25, no. 4, pp. 328–360, Apr. 2010, doi: https://doi.org/10.1108/02686901011034162.

[96] I. Khelil, H. Khlif, and I. Amara, "Political connections, political corruption and auditing: a literature review," *JFC*, vol. 29, no. 1, pp. 159–170, Jan. 2022, doi: 10.1108/JFC-12-2020-0257.

[97] S. Zahmatkesh and J. Rezazadeh, "The effect of auditor features on audit quality," *Tékhne*, vol. 15, no. 2, pp. 79–87, Jul. 2017, doi: https://doi.org/10.1016/j.tekhne.2017.09.003.

[98] S. W. Anderson, J. D. Daly, and M. F. Johnson, "Why Firms Seek ISO 9000 Certification: Regulatory Compliance or Competitive Advantage?," *Production and Operations Management*, vol. 8, no. 1, pp. 28–43, Jan. 2009, doi: https://doi.org/10.1111/j.1937-5956.1999.tb00059.x.

[99] M. Cameran, A. Ditillo, and A. Pettinicchio, "Audit Team Attributes Matter: How Diversity Affects Audit Quality," *European Accounting Review*, vol. 27, no. 4, pp. 595–621, Apr. 2017, doi: https://doi.org/10.1080/09638180.2017.1307131.

[100] X. Liu, B. Glocker, M. M. McCradden, M. Ghassemi, A. K. Denniston, and L. Oakden-Rayner, "The medical algorithmic audit," The Lancet Digital Health, vol. 4, no. 5, pp. e384–e397, May 2022, doi: 10.1016/S2589-7500(22)00003-6.

[101] V. Mahajan, V. K. Venugopal, M. Murugavel, and H. Mahajan, "The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms—How We Do It," Academic Radiology, vol. 27, no. 1, pp. 132–135, Jan. 2020, doi: 10.1016/j.acra.2019.09.009.

[102] S. Longpre et al., "A Safe Harbor for AI Evaluation and Red Teaming," 2024, arXiv. doi: 10.48550/ARXIV.2403.04893.

[103] N. Raemont, "Bank Customers Aren't Happy With AI Chatbots. Here's Why," *CNET*, Jun. 10, 2023. https://www.cnet.com/personal-finance/bank-customers-arent-happy-with-ai-chatbots-heres-why/

[104] T. Ryan-Mosley, "The movement to limit face recognition tech might finally get a win," *MIT Technology Review*, Jul. 20, 2023. https://www.technologyreview.com/2023/07/20/1076539/face-recognition-massachusetts-test-police/

[105] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[106] R. Gray, "Current Developments and Trends in Social and Environmental Auditing, Reporting and Attestation: A Review and Comment," *International Journal of Auditing*, vol. 4, no. 3, pp. 247–268, 2000, doi: 10.1111/1099-1123.00316.

[107] F. Delgado, S. Yang, M. Madaio, and Q. Yang, "The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice," presented at the EAAMO '23, Oct. 2023. doi: https://doi.org/10.1145/3617694.3623261.

[108] M. Feffer, M. Skirpan, Z. Lipton, and H. Heidari, "From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research," presented at the AIES '23, Aug. 2023. doi: https://doi.org/10.1145/3600211.3604661.

[109] S. Bergman, N. Marchal, J. Mellor, S. Mohamed, I. Gabriel, and W. Isaac, "STELA: a community-centred approach to norm elicitation for AI alignment," *Scientific Reports*, vol. 14, no. 1, Mar. 2024, doi: https://doi.org/10.1038/s41598-024-56648-4.

[110] M. K. Lee *et al.*, "WeBuildAI: Participatory Framework for Algorithmic Governance," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–35, Nov. 2019, doi: https://doi.org/10.1145/3359283.

[111] QueerInAI *et al.*, "Queer In AI: A Case Study in Community-Led Participatory AI," in *arXiv (Cornell University)*, Cornell University, Jun. 2023. doi: https://doi.org/10.1145/3593013.3594134.

[112] A, Mantelero, "The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template," *Computer Law & Security Review*, vol. 54, no. 106020, pp. 1–18, Sep. 2024, doi: https://doi.org/10.1016/j.clsr.2024.106020.

[113] Secretariat, Treasury Board of Canada, Algorithmic impact assessment tool. Treasury Board of Canada, 2021. [Online]. Available: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[114] O. Hankivsky *et al.*, "An intersectionality-based policy analysis framework: critical reflections on a methodology for advancing equity," *International Journal for Equity in Health*, vol. 13, no. 1, Dec. 2014, Available: https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-014-0119-x

[115] L. Gao, et al. "A framework for few-shot language model evaluation," Dec. 2023. Available: https://github.com/EleutherAI/lm-evaluation-harness

[116] Guidelines for auditing management systems, ISO Standard No. 19011: 2018, 2018. Available: https://www.iso.org/standard/70017.html

[117] P. Guldimann *et al.*, "COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act," *arXiv.org*, 2024. https://arxiv.org/abs/2410.07959 (accessed Nov. 20, 2024).

[118] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. Aroyo, "'Everyone wants to do the model work, not the data work': Data Cascades in High-Stakes AI," *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, doi: https://doi.org/10.1145/3411764.3445518.

[119] T. Tommasi, Novi, B. Caputo, and T. Tuytelaars, "A Deeper Look at Dataset Bias." Accessed: Aug. 06, 2024. [Online]. Available: https://arxiv.org/pdf/1505.01257

[120] G. Galdon Clavell, "AI Auditing - Checklist for AI Auditing," European Data Protection Board, 2023. Available: https://www.edpb.europa.eu/system/files/2024-06/ai-auditing_checklist-for-ai-auditing-scores_edpb-spe-programme_en.pdf

[121] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding Explainability: Towards Social Transparency in AI systems," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021, doi: https://doi.org/10.1145/3411764.3445188.

[122] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, and S. Kujala, "Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements," *Information and Software Technology*, vol. 159, no. 159, p. 107197, Mar. 2023, doi: https://doi.org/10.1016/j.infsof.2023.107197.

[123] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion*, vol. 58, no. 1, pp. 82–115, Jun. 2020.

[124] M. Nauta *et al.*, "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI," Feb. 2023, doi: https://doi.org/10.1145/3583558.

[125] A. Páez, "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)," *Minds and Machines*, vol. 29, no. 3, pp. 441–459, May 2019, doi: https://doi.org/10.1007/s11023-019-09502-w.

[126] L. Rice, E. Wong, and J. Zico Kolter , "Overfitting in adversarially robust deep learning," in *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria. Available: http://proceedings.mlr.press/v119/rice20a/rice20a.pdf

[127] Office of the Privacy Commissioner of Canada, "PIPEDA fair information principles," 2019. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/

[128] L. Gitelman, Ed., *"Raw Data" Is an Oxymoron*. The MIT Press, 2013.

[129] E. Rolf, T. Worledge, B. Recht, and M. I. Jordan, "Representation Matters: Assessing the Importance of Subgroup Allocations in Training Data," in *Proceedings of the 38th International Conference on Machine Learning*, 2021. Available: http://proceedings.mlr.press/v139/rolf21a/rolf21a.pdf

[130] A. Wang, V. V. Ramaswamy, and O. Russakovsky, "Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, Jun. 2022. doi: https://doi.org/10.1145/3531146.3533101.

[131] L. Taylor, "What Is Data justice? the Case for Connecting Digital Rights and Freedoms Globally," *Big Data & Society*, vol. 4, no. 2, pp. 1–14, Nov. 2017, doi: https://doi.org/10.1177/2053951717736335.

[132] I. Calzada, "Data Co-Operatives through Data Sovereignty," *Smart Cities*, vol. 4, no. 3, pp. 1158–1172, Sep. 2021, doi: https://doi.org/10.3390/smartcities4030062.

[133] T. Kukutai and J. Taylor, Eds., *Indigenous Data Sovereignty*. ANU Press, 2016. doi: https://doi.org/10.22459/caepr38.11.2016.

[134] R. Lovett, V. Lee, T. Kukutai, D. Cormack, S. C. Rainie, and J. Walker, "Good Data Practices for Indigenous Data Sovereignty and Governance," in *Good Data*, A. Daly, S. K. Devitt, and M. Mann, Eds., Amsterdam: Institute of Network Cultures, 2019, pp. 26–36. Available: https://networkcultures.org/blog/publication/tod-29-good-data/

[135] B. Neilson and N. Rossiter, "Automating Labour and the Spatial Politics of Data Centre Technologies," in *Topologies of Digital Work: How Digitalisation and Virtualisation Shape Working Spaces and Places*, M. Will-Zocholl and C. Roth-Ebner, Eds., Palgrave Macmillan Cham, 2022, pp. 77–101. Available: https://link.springer.com/chapter/10.1007/978-3-030-80327-8_4

[136] I. Ajunwa, K. Crawford, and J. Schultz, "Limitless Worker Surveillance," *California Law Review*, vol. 105, no. 3, pp. 735–776, 2017.

[137] L. Bui, "Asian Roboticism: Connecting Mechanized Labor to the Automation of Work," *Perspectives on Global Development and Technology*, vol. 19, no. 1–2, pp. 110–126, Mar. 2020, doi: https://doi.org/10.1163/15691497-12341544.

[138] Global Partnership on AI, "AI for Fair Work Report," Global Partnership on AI. Available: https://fair.work/wp-content/uploads/sites/17/2022/12/AI-for-fair-work-report-edited.pdf

[139] E. D. Williams, R. U. Ayres, and M. Heller, "The 1.7 Kilogram Microchip: Energy and Material Use in the Production of Semiconductor Devices," *Environmental Science & Technology*, vol. 36, no. 24, pp. 5504–5510, Dec. 2002, doi: https://doi.org/10.1021/es025643o.

[140] A. Villard, A. Lelah, and D. Brissaud, "Drawing a chip environmental profile: environmental indicators for the semiconductor industry," *Journal of Cleaner Production*, vol. 86, pp. 98–109, Jan. 2015, doi: https://doi.org/10.1016/j.jclepro.2014.08.061.

[141] F. Rohde *et al.*, "Broadening the perspective for sustainable artificial intelligence: sustainability criteria and indicators for Artificial Intelligence systems," *Current Opinion in Environmental Sustainability*, vol. 66, Feb. 2024, doi: https://doi.org/10.1016/j.cosust.2023.101411.

[142] S. Kallio and M. Farzan, "A transparent and standards-based way to assess the environmental impact of AI systems," *Nokia*, 2024. https://onestore.nokia.com/asset/214115 (accessed Oct. 12, 2024).

[143] B. Rakova and R. Dobbe, "Algorithms as Social-Ecological-Technological Systems: an Environmental Justice Lens on Algorithmic Audits," in 2023 ACM Conference on Fairness,

Accountability, and Transparency, Chicago IL USA: ACM, Jun. 2023, pp. 491–491. doi: 10.1145/3593013.3594014.

[144] R. Shelby et al., "Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction," in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, Montreal QC Canada: ACM, Aug. 2023, pp. 723–741. doi: 10.1145/3600211.3604673.

[145] A. Ebrahim and V. K. Rangan, "What Impact? A Framework for Measuring the Scale and Scope of Social Performance," *California Management Review*, vol. 56, no. 3, pp. 118–141, May 2014, doi: 10.1525/cmr.2014.56.3.118.

[146] J. Schuett, "Frontier AI developers need an internal audit function," *Risk Analysis*, Oct. 2024, doi: https://doi.org/10.1111/risa.17665.

[147] S. Casper et al., "Black-Box Access is Insufficient for Rigorous AI Audits," Jan. 25, 2024, arXiv: arXiv:2401.14446. Accessed: Feb. 29, 2024. [Online]. Available: http://arxiv.org/abs/2401.14446

[148] OpenAI, "GPT-4o System Card," 2024. https://openai.com/index/gpt-4o-system-card/

[149] Meta, "Llama 3.1 Model Card," 2024. GitHub repository, https://github.com/meta-llama/llama-models/commits/main/models/llama3_1/MODEL_CARD.md

[150] C. Chouhan, C. M. LaPerriere, Z. Aljallad, J. Kropczynski, H. Lipford, and P. J. Wisniewski, "Co-designing for Community Oversight," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–31, Nov. 2019, doi: https://doi.org/10.1145/3359248.

[151] F. Skopik, G. Settanni, and R. Fiedler, "A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing," *Computers & Security*, vol. 60, pp. 154–176, Jul. 2016, doi: https://doi.org/10.1016/j.cose.2016.04.003.

[152] A. Bussone *et al.*, "Trust, Identity, Privacy, and Security Considerations for Designing a Peer Data Sharing Platform Between People Living With HIV," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–27, Oct. 2020, doi: https://doi.org/10.1145/3415244.

[153] Y. Chen and H. Xu, "Privacy management in dynamic groups: understanding information privacy in medical practices," *CSCW '13: Proceedings of the 2013 conference on Computer supported cooperative work*, Feb. 2013, doi: https://doi.org/10.1145/2441776.2441837.

[154] T. Tiropanis, W. Hall, J. Hendler, and C. de Larrinaga, "The Web Observatory: A Middle Layer for Broad Data," *Big Data*, vol. 2, no. 3, pp. 129–133, Sep. 2014, doi: https://doi.org/10.1089/big.2014.0035.

[155] N. Hamed, O. Rana, P. Orozco-terWengel, B. Goossens, and C. Perera, "A Comparison of Open Data Observatories," *ACM Journal of Data and Information Quality*, Nov. 2024, doi: https://doi.org/10.1145/3705863.

[156] A. Miceli, A. Dinika, K. Kauffman, C. Salim Wagner, and L. Sachenbacher. *The Data Workers' Inquiry.* https://data-workers.org

[157] P. M. Krafft et al., "An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists," in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada: ACM, Mar. 2021, pp. 772–781. doi: 10.1145/3442188.3445938.

[158] M. S. Lam, A. Pandit, C. H. Kalicki, R. Gupta, P. Sahoo, and D. Metaxa, "Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising," Proc. ACM Hum.-Comput. Interact., vol. 7, no. CSCW2, pp. 1–37, Sep. 2023, doi: 10.1145/3610209.

[159] M. Amirizaniani, T. Roosta, A. Chadha, and C. Shah, "AuditLLM: A Tool for Auditing Large Language Models Using Multiprobe Approach," Feb. 14, 2024, arXiv: arXiv:2402.09334. [Online]. Available: http://arxiv.org/abs/2402.09334

[160] J. Bandy and N. Diakopoulos, "Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News," Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 36–47, May 2020, doi: 10.1609/icwsm.v14i1.7277

[161] E. Bokányi and A. Hannák, "Understanding Inequalities in Ride-Hailing Services Through Simulations," Sci Rep, vol. 10, no. 1, p. 6500, Apr. 2020, doi: 10.1038/s41598-020-63171-9

[162] E. McCullough, B. Dolber, J. Scoggins, E.-M. Muña, and S. Treuhaft, "Prop 22 Depresses Wages and Deepens Inequities for California Workers," *nationalequityatlas.org*, 2022. https://nationalequityatlas.org/prop22-paystudy

[163] P. Adler et al., "Auditing black-box models for indirect influence," Knowl Inf Syst, vol. 54, no. 1, Art. no. 1, Jan. 2018, doi: 10.1007/s10115-017-1116-3

[164] J. Hernández-Orallo, "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement," *Artificial Intelligence Review*, vol. 48, no. 3, pp. 397–447, Aug. 2016, doi: https://doi.org/10.1007/s10462-016-9505-7.

[165] P. Vaithilingam, T. Zhang, and E. L. Glassman, "Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, Apr. 2022. doi: https://doi.org/10.1145/3491101.3519665.

[166] Linux Foundation AI & Data Foundation, "Adversarial Robustness Toolbox (ART)", 2019. GitHub repository, https://github.com/Trusted-AI/adversarial-robustness-toolbox.

[167] C. T. Wolf, "Explainability scenarios: towards scenario-based XAI design," *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Mar. 2019, doi: https://doi.org/10.1145/3301275.3302317.

[168] D. Fremont *et al.*, "Formal Scenario-Based Testing of Autonomous Vehicles: From Simulation to the Real World," 2020. Available: https://arxiv.org/pdf/2003.07739

[169] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

[170] M. S. Lam, M. L. Gordon, D. Metaxa, J. T. Hancock, J. A. Landay, and M. S. Bernstein, "End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior," Proc. ACM Hum.-Comput. Interact., vol. 6, no. CSCW2, pp. 1–34, Nov. 2022, doi: 10.1145/3555625.

[171] M. Sloane, E. Moss, O. Awomolo, and L. Forlano, "Participation Is not a Design Fix for Machine Learning," *Equity and Access in Algorithms, Mechanisms, and Optimization*, Oct. 2022, doi: https://doi.org/10.1145/3551624.3555285.

[172] E. A. Coleman, J. Manyindo, A. R. Parker, and B. Schultz, "Stakeholder engagement increases transparency, satisfaction, and civic action," *Proceedings of the National Academy of Sciences*, vol. 116, no. 49, pp. 24486–24491, Nov. 2019, doi: https://doi.org/10.1073/pnas.1908433116.

[173] The International Association of Algorithmic Auditors, "AI Auditing Definitions," 2024. Available: https://iaaa-algorithmicauditors.org/wp-content/uploads/2024/10/IAAA-AI-Auditing-Definitions-1.pdf

[174] V. Ojewale, R. Steed, B. Vecchione, A. Birhane, and I. D. Raji, "Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling," Mar. 14, 2024, *arXiv*: arXiv:2402.17861. Accessed: Nov. 21, 2024. [Online]. Available: http://arxiv.org/abs/2402.17861

[175] S. Cattell, A. Ghosh, and L.-A. Kaffee, "Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities," *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24)*, vol. 7, pp. 267–280, Oct. 2024, doi: https://doi.org/10.1609/aies.v7i1.31635.

[176] I. D. Raji and J. Buolamwini, “Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products,” in Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu HI USA: ACM, Jan. 2019, pp. 429–435. doi: 10.1145/3306618.3314244.

[177] M. Raymond and L. DeNardis, “Multistakeholderism: anatomy of an inchoate global institution,” *Int. Theory*, vol. 7, no. 3, pp. 572–616, Nov. 2015, doi: 10.1017/S1752971915000081.

ACKNOWLEDGMENTS

Contributors

Drafting authors: Benjamin Faveri (Consulting Scientist, Canada), Maureen Johnson-León (Consulting Scientist, USA), Prem Sylvester (Consulting Scientist and Panel Manager, India/Canada), Wendy Hui Kyong Chun (Panel Chair, Canada), Anikó Hannák (Panel Vice-Chair, Switzerland/Hungary), Marcelo Mendoza (Panel Vice-Chair, Chile), Meredith Broussard (Member, United States), Ruben Enikolopov (Member, Spain/Russia), Alondra Nelson (Member, United States), Christian Sandvig (Member, United States), ‘Gbenga Sesan (Member, Nigeria), Andrew Sporle (Member, New Zealand), Rohini Srihari (Member, United States), Janaki Srinivasan (Member, India/United Kingdom), Christo Wilson (Member, United States), Mimi Zou (Member, United Kingdom/Australia). Contributors: Sebastián Valenzuela (IPIE Chief Science Officer, Chile), Philip Howard (IPIE President and CEO, Canada/United Kingdom). Copyediting: Romilly Golding and Bervely Skykes. We gratefully acknowledge support from the IPIE Secretariat: Egerton Neto, Donna Seymour, and Alex Young. We thank the following experts for providing testimony to the Panel: Taka Ariga (United States), Stephanie Bell (United States), Olivia J. Erdelyi (Germany/Hungary/New Zealand), Grace Githaiga (Kenya), Gemma Galdon Clavell (United States/Spain), Vidushi Marda (India), Anna-Katharina Meßmer (Germany/Europe), Jakob Mökander (United Kingdom), Irene Mwendwa (Kenya/Uganda), Nayantara Ranganathan (India), Briana Vecchione (United States), Leon Yin (United States). We also acknowledge feedback provided by IPIE affiliates: Luis Adrián Castro-Quiroa, Eloísa García-Canseco, Joan Tirwyn Hassan, Piers Howe, Ozan Kuru, Anita Lavorgna, Wenlong Li, Nihat Muğurtay, Fredrick Ogenga, Dariia Opryshko, Tanja Pavleska, Margarita Sordo, Jun Zhang.

Funders

The International Panel on the Information Environment (IPIE) gratefully acknowledges the support of the Children’s Investment Fund Foundation, Ford Foundation, Heising-Simons Foundation, Oak Foundation, and the Simons Foundation. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the IPIE and do not necessarily reflect the views of the funders. For an updated list of funding partners, please check www.IPIE.info.

Declaration of Interests

IPIE reports are developed and reviewed by a global network of research affiliates and consulting scientists who constitute focused Scientific Panels and contributor teams. All contributors and reviewers complete declarations of interests, which are reviewed by the IPIE at the appropriate stages of work.

Preferred Citation

An IPIE *Summary for Policymakers* provides a high-level state of the art and is written for a broad audience. An IPIE *Synthesis Report* makes use of scientific meta-analysis techniques, systematic review, and other tools for evidence aggregation, knowledge generalization, and

scientific consensus building, and is written for an expert audience. An IPIE *Technical Report* addresses particular questions of methodology, or provides a policy analysis on a focused regulatory problem. All reports are available on the IPIE website (www.IPIE.info). This document should be cited as:

International Panel on the Information Environment, "Towards A Global AI Auditing Framework: Assessment and Recommendations" Zurich, Switzerland, SR2024.3, 2024.

DOI
https://IPIE.info/10.61452/ZWED1485

Copyright Information

ABOUT THE IPIE

The International Panel on the Information Environment (IPIE) is an independent and global science organization providing scientific knowledge about the health of the world's information environment. Based in Switzerland, the IPIE offers policymakers, industry, and civil society actionable scientific assessments about threats to the information environment, including AI bias, algorithmic manipulation, and disinformation. The IPIE is the only scientific body systematically organizing, evaluating, and elevating research with the broad aim of improving the global information environment. Hundreds of researchers worldwide contribute to the IPIE's reports.

For more information, please contact the International Panel on the Information Environment (IPIE), secretariat@IPIE.info. Seefeldstrasse 123, P.O. Box, 8034 Zurich, Switzerland.

International Panel on the Information Environment

Seefeldstrasse 123
P.O. Box 8034 Zurich
Switzerland

ISBN 978-3-03983-005-3