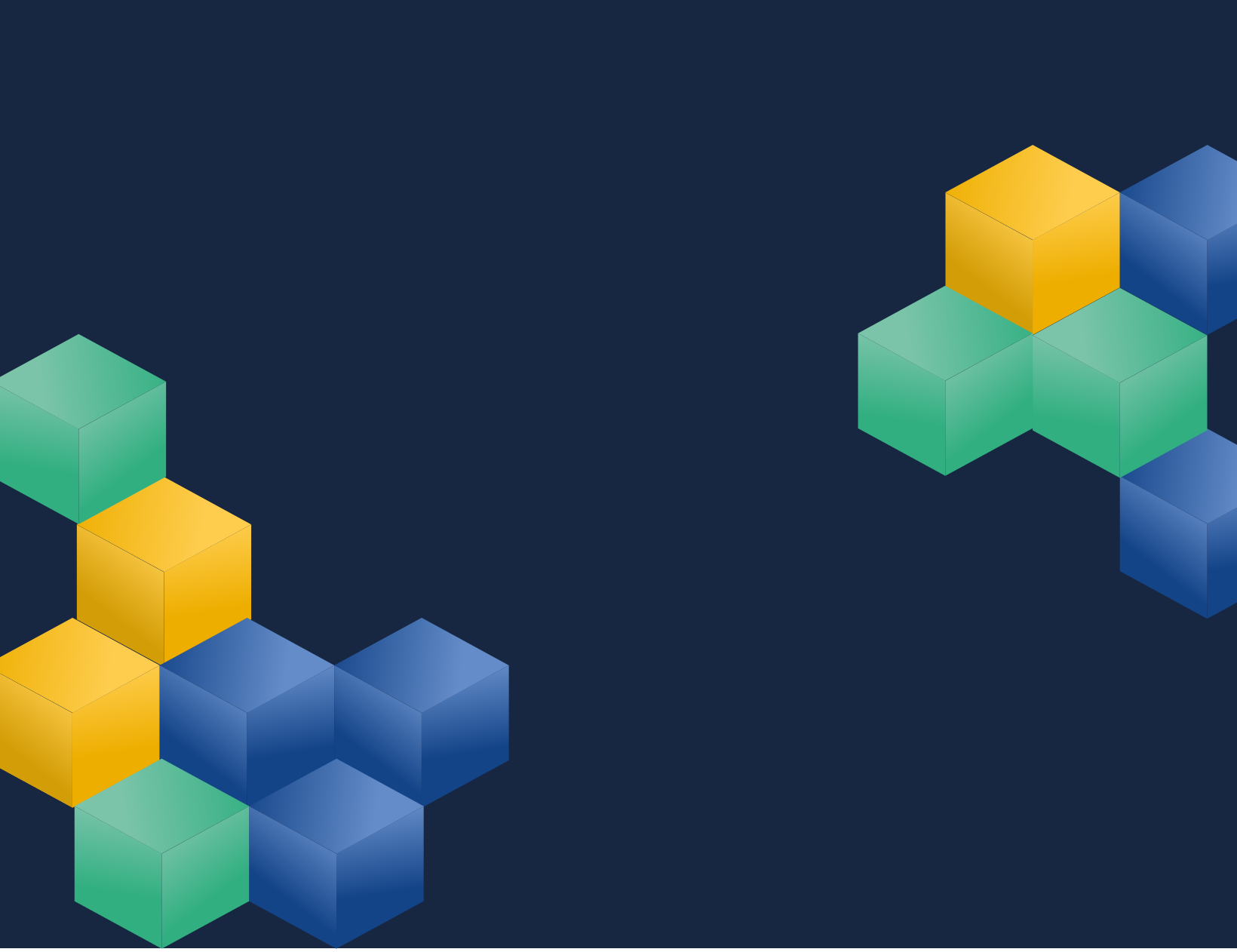


IPIE
International Panel on the Information Environment

Global Approaches to Auditing Artificial Intelligence

A Literature Review

Synthesis Report 2024.1



IPIE
International Panel on the Information Environment

Global Approaches to Auditing Artificial Intelligence

A Literature Review

Synthesis Report 2024.1

SYNOPSIS

The risks and opportunities of AI systems have meant that governments, companies, and civil society actors across the world are turning to AI audits to evaluate these systems. Audits offer a burgeoning set of means to test whether algorithmic or AI systems engender the outcomes they are expected to or whether they have significant – possibly adverse – societal and technological impacts.

This *Synthesis Report* (SR2024.1) is a literature review outlining the regulatory, industry, and academic approaches to AI audits. We review 78 articles published in peer-reviewed journals and as preprints, 21 documents from industry associations and standard-setting organizations, and national policy documents and regulations from 20 countries.

Based on this review, we identify three key takeaways about the landscape of AI auditing:

1. To accurately assess the potential risks, and impacts of AI systems, we need a trustworthy audit ecosystem with complementary approaches from internal, external, and community auditors.

2. Auditors need better access to data and audit artifacts from the developers and deployers of AI systems. Comprehensive auditability requires documentation and disclosure of an AI system's model and data components, associated risks and impacts, and easily understandable explanations of its outcomes.

3. Given that the development and use of AI systems impacts communities across the world, audit regimes must account for their global effects. Most existing audits have been conducted in North America, Europe, and other regions of the 'global north', with their results typically published in English and focused on effects within these regions. The impacts of AI systems, however, include shifts in social and environmental conditions beyond the immediate development or application contexts of a system.

Given the diverse audit ecosystem we find, we recommend that global standards for AI audits are required to build trust in the quality and rigor of different audits. This review therefore aims to inform the development of criteria, evaluation mechanisms, evidentiary documentation, and processes crucial to such standards.

CONTENTS

SYNOPSIS **3**

SECTION 1. INTRODUCTION **5**

WHAT ARE ARTIFICIAL INTELLIGENCE (AI) AUDITS, AND WHY DO THEY MATTER? 5

SECTION 2. APPROACHES TO AI AUDITS **7**

WHAT APPROACHES TO AI AUDITS ARE PRESENT IN VARIOUS REGULATORY, INDUSTRY, AND ACADEMIC FRAMEWORKS? 7

Auditing AI systems for risk and conformity *8*

Institutional approaches to AI auditing 9

Enforcement mechanisms for AI audits 11

Private industries' approaches to AI auditing 15

Documentation for AI audits as governance mechanisms 16

Auditing the performance of AI systems *17*

Who 19

What and when 22

Why 25

How 27

What next 30

SECTION 3. DISCUSSION AND CONCLUSION **34**

ENDNOTES **38**

REFERENCES **39**

APPENDIX A: A NOTE ON METHODOLOGY **51**

APPENDIX B: NATIONAL AI POLICIES THAT RECOMMEND AUDITS **53**

ACKNOWLEDGMENTS **56**

CONTRIBUTORS 56

FUNDERS 56

DECLARATION OF INTERESTS 56

PREFERRED CITATION 57

DOI 57

COPYRIGHT INFORMATION 57

ABOUT THE IPIE **58**

SECTION 1. INTRODUCTION

What are Artificial Intelligence (AI) Audits, and Why do They Matter?

The risks and opportunities of AI systems[1], as well as their inscrutability (if not opacity), have meant that governments, institutions, and standard-setting bodies across the world are seeking ways to evaluate whether and how these systems function. These actors have therefore turned to AI (or algorithm) audits.

Audits offer a burgeoning set of means to test whether algorithmic or AI systems engender the outcomes they are expected to and/or whether they have significant – possibly adverse – societal and technological impacts. These audits probe an AI system's design, development, and operations, often examining the model and data used in it. They are used to describe how the audited AI system performs against certain established criteria and to report on its impacts. Audits are seen as mechanisms for assessing AI systems' alignment with norms and principles of AI responsibility, accountability, trustworthiness, or safety [1], [2], [3], [4]. The diversity of actors, approaches, and purposes of audits, however, has produced a fragmented ecosystem of AI audits, making it difficult to assess the global impacts of these systems. Frameworks, guidelines, and standards for audits are being developed under the rubric of AI governance. Private actors, such as consultancies and technology companies, have also begun developing methods and frameworks for these types of audits that are meant to demonstrate compliance with legal, regulatory, ethical, sectoral, and/or organizational policies and principles. The increasing risks of adverse impacts that AI systems pose in organizational, institutional, and societal contexts have also contributed to the growing interest in AI audits. These audits frequently assess the performance of AI systems — the extent to which they do or do not align with a set of established criteria — based on a set of questions that orient the aims, methods, and outcomes of the audit.

Often, audits assess AI systems for their divergence from certain established performance criteria concerned with societal wellbeing. Such criteria typically include bias or discrimination along axes of race, gender, age, socioeconomic position, or other demographic categories [5], [6]. Audits may also consider criteria of ethical fairness [7], [8]. Some researchers have called for AI audits to be an activist task that can mitigate the disproportionate impacts or harms these technologies, if used carelessly, can have on historically marginalized peoples [9]. Audits therefore help align the development and use of AI systems with socially beneficial outcomes.

This literature review outlines the regulatory, industry, and academic approaches to AI audits. There is, however, a definitional ambiguity in referring to an audit of certain algorithmic systems as AI, algorithm, or algorithmic audits. Algorithm audits were first proposed to assess the outputs of algorithmic systems [10]. Algorithmic audits used to refer to forms of auditing where algorithmic systems are taken not as the object of audits but instead used as decision aids. AI audits share the concerns of the other forms of audits but typically take as their object data-driven AI systems that use machine-learning models. However, such differentiation is currently less clear. AI audits have come to describe various audits of algorithmic systems (including AI systems) that also consider the stakeholder risks and impacts as well as auditee conformity with organizational and public policy. These audits have blurred disciplinary and programmatic genealogies and concerns. We have retained the terminology of AI audits to imply the diversity of these approaches "not as a methodological problem but a substantive fact" [11, p. 6]. We review 78 articles published in peer-reviewed journals and as preprints, 21 documents from industry associations and standard-setting organizations, and national policy documents and regulation from 20 countries (see Appendix A for an extended discussion of methodology). Given the diverse audit ecosystem we find, standards can build trust in the quality and rigor of different audits. This review therefore aims to inform the

development of criteria, evaluation mechanisms, evidentiary documentation, and processes crucial to global standards for AI audits.

SECTION 2. APPROACHES TO AI AUDITS

What Approaches to AI Audits are Present in Various Regulatory, Industry, and Academic Frameworks?

Frameworks and standards for AI audits delineate the principles, processes, and procedures relevant to auditing AI systems. Whitepapers, standards, and other advisory documents drafted by regulatory and industry bodies such as the UK's Information Commissioner's Office (ICO), the U.S.-based Government Accountability Office (GAO) and National Institute of Standards and Technology (NIST), Singapore's Infocomm Media Development Authority (IMDA), and the Institute of Internal Auditors (IIA) are broadly concerned with organizational or institutional AI governance. Such mechanisms typically stem from the history of operational and compliance audits (conducted internally or externally to an organization) that assess an organization's regular operations, planning, and processes, according to internal policies and regulations. These AI audits also draw on social and safety audits that assess an entity or system's alignment with social wellbeing and public safety. First attempts at enforcement regimes that incorporate AI audits have been made via the EU's AI Act of 2024 and New York City's Local Law 144 in 2021. Whereas audits in the regulatory, industrial, or organizational case compare AI systems' operations and outcomes with a certain set of established policies and standards, academic approaches typically test AI systems for the direct or indirect impacts they might have on individuals or members of the general public. Such audits are motivated to test AI systems for specific instances of bias, discrimination, unfairness or other

forms of societal harms, which is in line with the social-scientific history of audit studies [9], [10].

Auditing AI systems for risk and conformity

Audits in organizational and institutional contexts are often used to evaluate an organization's management and governance of AI systems in terms of risk and conformity assessments.

As risk assessments, audits evaluate whether the use of an AI system, including its data and algorithmic components, risks having an adverse impact on stakeholders and organizations. According to the U.S. National Institute of Standards and Technology (NIST) AI Risk Management Framework (RMF), "risk is a function of 1) [an AI system's] negative impact, or magnitude of harm [to people, organizations, and/or ecosystems], that would arise [from the system's operation] and 2) the likelihood of [that impact's or harm's] occurrence" [3, p. 5]. These frameworks situate the risk levels of AI systems on a spectrum (such as the AI RMF or the Canadian Treasury Board's Algorithmic Impact Assessments) or against a threshold of high risk or high impact (such as in the EU AI Act). Some frameworks leave the enumeration of risk to the organizations themselves. Within such risk-centered frameworks, bias, discrimination, and other ethical concerns that might lead companies to incur reputational or financial risks are relevant.

When an audit assesses whether the organizational structures, processes, and policies that govern and manage the use of AI systems conform to certain standards, it is more accurately described as a conformity assessment. In line with traditional internal audits, this form of audit considers risk and impact assessment as supplementary to the audit, rather than the goal.

Audits (such as in the EU AI Act or the ISO/IEC 42001:2023 standard), assess an organization's conformity with regulation and international standards alongside mechanisms to manage the risks associated with developing or deploying an AI system. Such audits assess the entire life cycle of an AI system, from design to deployment. The audit criteria include an organization's internal controls developed concurrently with their AI policy. Audits also assess internal controls instituted to manage risks from third parties involved in any stage of the AI system's life cycle, such as component suppliers and customers. For private organizations, planning, developing, and reviewing their conformity with such policies typically involves extending existing structures and governance and risk management checks [13]. Such extensions may include:

- assigning roles and responsibilities for the oversight of an AI system and its various components across the system's life cycle;
- implementing appropriate controls to check for misalignments with organizational policy and legal and regulatory compliance;
- monitoring the performance of AI systems, such as for bias and ethics compliance.

AI audits function as risk or conformity assessments in various organizational, institutional, and legal contexts.

Institutional approaches to AI auditing

Guidelines for AI audits have often come from professional auditor bodies typically located in the U.S.A., although they claim an international membership. These recommendations generally extend existing auditing frameworks for organizational governance, such as Control Objectives for Information Technology (COBIT) 2019, created by the Information Systems Audit and Control Association (ISACA) [12]. This scope of auditing is rapidly evolving to account specifically for AI systems' human

and technical components. In the Institute of Internal Auditors' (IIA's) AI Auditing Framework, the "human factor" pertains to operating and using the AI system legally, ethically, and responsibly, as well as making the "black box" of these systems legible to organizational stakeholders [13]. At the technical end, this framework also recommends identifying algorithmic bias and auditing the organization's checks on data quality and infrastructure.

Institutional frameworks for AI use are meant to demonstrate the institution's accountability and trustworthiness, as well as its risk management capacities. These are largely developed for the U.S.A., and certain national territories in Western countries, although sometimes they are cited by entities in other jurisdictions. The U.S. Government Accountability Office's AI accountability framework for governmental institutions recommends that government entities evaluate AI systems at both the component (or model) and the systemic levels [2]. This framework lists the questions auditors may ask when conducting such an assessment of AI systems, along with specific audit procedures for obtaining the necessary information. Furthermore, it recommends that auditors engage with both "AI technical stakeholders—data scientists, data engineers, developers, cybersecurity specialists, program managers" and "a broader community of stakeholders—policy and legal experts, subject matter experts, [civil liberties advocates], and individuals using the AI system or impacted by its use, among others" [2, p. 18-19]. Other frameworks that introduce an international lens provide recommendations for non-governmental sector-specific assessment of AI systems, such as the framework developed for healthcare by the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) [14].

These frameworks are all concerned with the human components of AI systems—including an organization's executives and technical staff, and external stakeholders such as affected communities—and their interactions with technical dimensions,

such as the models and training data. They recommend audits to assess whether organizations and institutions adhere to socially beneficial practices when using AI systems. These audits might comprise interviews with internal and external stakeholders, access to internal documentation and information about quality controls, and technical tests to check for algorithmic biases and dataset quality. These audits surface different ways of ensuring rigor and identifying potential laxities in the design and operation of AI systems.

National and international standard-setting organizations are also working towards creating standards for auditing AI systems. For example, in the U.S., the Artificial Intelligence Risk Management Framework (AI RMF 1.0) from the National Institute for Standards and Technology (NIST) recommends that auditors participate in the operations and monitoring stages of an AI system's deployment to ensure that it is trustworthy. Audits determine trustworthiness by assessing whether a system is "valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed" [3, p. 3]. The ISO/IEC 42001:2023 standard recommends internal audits as a key tool for evaluating the technical performance and organizational management of AI systems. These audits are primarily concerned with how AI systems align with organizational policy, and whether these systems include responsible design and development [15].

Enforcement mechanisms for AI audits

Although guidelines and frameworks for AI governance recommend audits to evaluate the development, deployment, and use of AI systems (see Appendix B for a list of nonbinding national approaches to AI governance that employ audits), only the U.S.A., Canada, the E.U., the U.K. and China have enforceable regulations that recommend or mandate the use of AI audits for private and governmental entities to

assess whether AI systems comply with legal and social norms and values. They differ in their classification of AI systems, the specificity of their recommendations regarding auditing procedures, criteria, and/or evidence, which aspects of these systems are subject to scrutiny, and the extent of that scrutiny.

Some legislation recommends governance mechanisms that test how AI systems align with principles considered socially consequential, such as transparency, accountability, oversight, accuracy, and documentation. However, such legislation does not prescribe the methodologies that should be used to assess such alignment.

The E.U.'s Digital Services Act (DSA) and Online Safety Act (OSA) carve out specific roles for internal and external auditors, as well as some auditor-like actors from academia and civil society, to assess the algorithms used by Very Large Online Platforms and Search Engines (VLOPs/VLOSEs) [16]. These audits cover a specific set of algorithms that may include—but do not account for all the uses of—AI systems.

The E.U.'s AI Act (AIA) goes further by identifying AI systems that pose a high risk to public safety, vulnerable populations, or critical social infrastructure. The AIA outlines conformity assessments and post-market monitoring plans that some argue can be construed as (internal or external) audits [17]. Conformity assessments may be conducted internally based on controls and related checks established within an organization or by external entities, such as notified bodies like the European Artificial Intelligence Board or national supervisory authorities established by the AIA. Similar to continuous auditing, post-market monitoring requires that developers and deployers of AI systems continuously assess high-risk AI systems for potential impacts. The use of high-risk AI systems must also be registered in a central database that is much like China's Algorithm Registry although the latter does not differentiate systems by their level of risk [18]. While the AIA outlines the documentation that an

organization must produce to meet the conformity assessment's requirements, it does not describe the methods to be used to audit an AI system. In Italy, draft AI regulation would task the Agency for Digital Italy (AgID) with evaluation and monitoring duties, such as accrediting independent auditors responsible for the protection of privacy, public safety, and consumer protection.

Similarly, Canada recommends audits as a means of oversight for high-impact AI systems and for holding generative AI systems accountable. The proposed Artificial Intelligence and Data Act (AIDA) emphasizes independent audits for AI systems that are particularly prone to harm. The supplementary Code of Practice or "guardrails" (for ensuring voluntary regulatory compliance for generative AI system developers, deployers, and operators) recommends both internal and external audits for any such system [19]. The algorithm and data that comprise an automated system are central concerns for the Algorithmic Impact Assessments mandated by the Canadian Treasury Board's Directive on Automated Decision Making [20]. These assessments also require that organizations assess how they consult with internal and external stakeholders about an AI system's impacts to adequately mitigate its adverse effects. However, the impact assessment does not require independent evaluation of the AI system.

When comparing the European and Canadian frameworks, the distinction between risk and impact is unclear, although they might coincide (such as in biometric systems and autonomous vehicles). Both frameworks and their governance structures emphasize the criteria for compliance, or conformity, rather than the procedures used for those evaluations.

While the U.S.A. currently has no federal regime in place for governing AI, some local regulations, such as New York City's Local Law 144 regarding automated employment decision tools (AEDTs), invoke audits as an assessment tool. This law

"prohibits employers and employment agencies from using an automated employment decision tool unless [it] has been subject to a bias audit" and its results are publicly available [21, p. 1]. In contrast to EU and Canadian frameworks discussed previously, this law presents fine-grained criteria and procedures for audits.

These audits rely on an "impact ratio" that measures AEDTs' differential impact on various demographic groups. The law also contains provisions regarding how an AEDT provider or user must respond if that ratio is skewed in a discriminatory manner against certain groups. Such audits identify specific socially sensitive scenarios in which the design or deployment of AI systems can harm vulnerable people. In the U.K., audit frameworks currently focus on regulatory concerns with specific aspects of AI systems, such as data. The Information Commissioner's Office (ICO) framework for auditing AI assesses whether the organization has implemented best practices for data protection compliance and appropriate safeguards for the development and deployment of AI systems. Such an audit assesses "the organisation's procedures, processes, records and activities" to ensure that "data [is] being used fairly, lawfully and transparently, [and that] people understand how their data is being used and how is data being kept secure" [22, p. 3]. These audits may also evaluate AI systems for statistical accuracy, discrimination, and bias in their data.

The reviewed regulatory frameworks offer diverse—often competing—tools and methods for assessing normative and regulatory compliance. The E.U.'s AIA and Canada's AIDA also foster an AI auditing 'ecosystem' that brings together auditors, regulatory bodies, and private companies [17]. Several existing audit firms, as well as newer ones dedicated to AI audits, have turned their attention to the demands of AI governance.

Private industries' approaches to AI auditing

The largest global audit firms, such as Deloitte, KPMG, PwC, and EY have incorporated mechanisms for assessing and ensuring compliance with AI governance regimes and norms as part of their audit services [23], [24], [25], [26]. Major consultancies such as McKinsey and BCG also offer custom-built tools for organizations to implement "ethical AI" or principles of "responsible AI" [27], [28].

Major technology companies, such as Meta, whose Responsible AI team was disbanded in 2023, as well as Google, Amazon, and OpenAI, have also published principles of safety, trust, security, and responsibility to which they claim their AI systems adhere [29], [30], [31], [32]. These companies sometimes claim they use audit-like tools and methods internally to evaluate this adherence. Most do not mention specific tools, although Meta describes its use of Robust Bias Evaluation of Large Generative Language Models (ROBBIE), "a tool that compares six different prompt-based bias and toxicity metrics across 12 demographic axes and five different LLMs" and DIG In which "focuses on evaluating gaps in quality and diversity of content generated from text-to-image models between geographic regions" [33].

Increasingly, there are dedicated AI audit firms, who have published about their audit instruments. Employees of Holistic AI, for example, have published surveys on algorithm auditing that are used to guide their audit services' methodology [34]. Other firms have published their own methodologies for auditing the ethical commitments of AI systems, such as BABL AI's "audit instrument" and ORCAA's "fairness matrix"[7], [35]. Eticas AI has also published the results of one of its "algorithmic audits" and an outline of its methodology [36]. These audits are typically aligned with existing legal regimes, such as New York City's Local Law 144 (see previous section). Notably, these publications offer little detail about specific audits they have conducted.

Documentation for AI audits as governance mechanisms

As the previous sections show, regulatory, organizational and private audits are increasingly aligned in their goals of assessing how AI systems are managed by, interact with, and impact various stakeholders. Such audits rely on the documentation of organizations' computational, data, and human resources.

The ISO/IEC 42001:2023 standard for organizations' management of AI systems recommends how data resources should be documented (including the provenance of the data, the associated labeling processes, data quality, any biases, and the preparation processes). The standard also mandates documentation of the algorithms and models employed in the system, model optimization and evaluation methods, and general computing resources such as network usage and data storage. Such documentation and related information required to understand and assess the system's risks must be available to users and any interested external parties interested in adverse impacts of the system. Audits for conformity with this standard should assess whether the human resources involved in the system, such as data scientists, experts on "trustworthiness" topics such as safety, security, privacy, and human oversight, and AI domain experts, adhere to organizational AI policy.

National bodies have also produced recommendations for AI system documentation. Singapore's IMDA's Model AI Governance Framework focuses on traceability and auditability to account for "the intimate interaction between data and algorithm/model" [38, p. 36]. Traceability requires that easily understood documentation (such as an audit trail or log) about an AI system's decisions and the data and processes that feed into it be made readily available to auditors. The AI system's auditability "refers to the [system's] readiness to undergo an assessment of its algorithms, data and design processes" [38, p. 51]. Similarly, the Canadian Treasury Board mandates documentation and evaluation of algorithmic and AI

systems used by federal bodies through its Algorithmic Impact Assessment Tool [20]. The framework for AI accountability from the U.S. Government Accountability Office also requires federal institutions document their use of AI systems for audits [2].

Many of the organizational and institutional requirements around AI auditing center on the assessment of governance structures, oversight processes, and organizational policies in the context of data-driven algorithmic, or AI, systems. As such, they enroll concerns around algorithmic bias, model design, and data documentation. These audits are based on evidence that demonstrates organizational alignment or regulatory compliance. Frequently, they are extensions or expansions of existing auditing frameworks used by internal auditors, financial auditors, or other forms of organizational auditors. But the kind of algorithm, or AI, audits that have garnered interest in the academic community are more accurately located in the history of social-scientific audit studies. They intend to draw attention to how AI systems work—in theory and in practice—and the bias, discrimination, or other forms of societal harms that they may engender.

Auditing the performance of AI systems

The approaches to AI audits suggested in auditing frameworks and standards offer important ways for organizations, institutions, and regulators to assess AI systems and their interactions with various stakeholders. These stakeholders can range from an organization developing AI systems to society at large. However, these frameworks are less clear about how those potential impacts, especially in the form of social harms, might be described and measured.

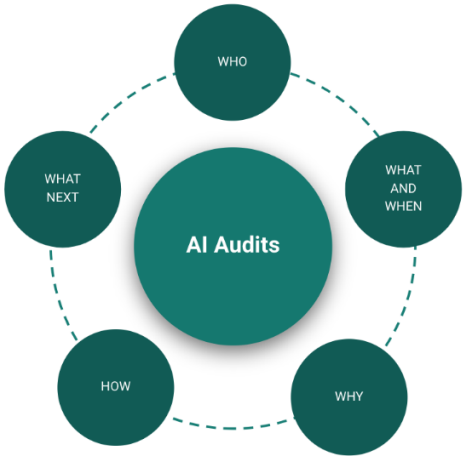
Some frameworks for AI audits have been suggested in the academic literature to engage with some of the concerns of AI governance-oriented audits. The Scoping, Mapping, Artifact Collection, Testing and Reflection (SMACTR) framework proposed by Raji et al. for internal audits is meant to be operationalized “end to end”

throughout the life cycle of an organization's AI system development [38]. The Generalized Audit Framework for AI (GAFAI) adopts a "technical oriented approach with measurable objectives as indication for a system's trustworthiness and considers the entire AI lifecycle" [39, p. 1,248]. Oala et al. demonstrate ways to translate the Machine Learning for Health (ML4H) audit framework proposed by the ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) into practice [14].

These AI audits largely evaluate the performance of AI systems against their intended or expected operation, especially to discover societal harms or other forms of "problematic machine behaviour" emerging from those systems [40]. Such performance evaluation describes whether AI systems operate in line with certain predetermined criteria or metrics, or to what degree their performance deviates from an acceptable baseline. Those performance criteria can be established by the auditor in collaboration with the company developing or using the AI system, or in law [41], [42], [43].

AI audits frequently assess how the operations of AI systems result in biases against certain people, especially individuals belonging to historically marginalized or legally protected demographic groups. Audits also assess AI systems for the possibility of widespread social impacts, such as the loss of privacy or the spread of disinformation. However, the literature does not suggest standard methodologies or tools for conducting AI audits; they vary across auditors and audit targets. This diversity becomes especially consequential when auditing Large Language Models (LLMs) and other contemporary AI systems, such as foundational models and multimodal) generative AI. It is useful, then, to consider the interrelated questions that guide various AI audits in the academic literature (Figure 1).

Figure 1. Questions Guiding AI Audits.



Source: IPIE analysis based on this literature review.

Who

Auditors have diverse relations to the companies and organizations whose AI systems they audit. Each auditor brings different expertise and varying levels of access to the systems being audited. The individuals or entities that conduct audits can be described as follows:

- **Internal auditors**: As has been the case with organizational audits, the company has an internal auditing team specifically for AI systems, with associated roles, responsibilities, processes, and procedures. Scholars recommend the involvement of such a team in AI audits as it enables a significant degree of access to the development, deployment, and maintenance of AI systems [34], [38]. These auditors aim to understand whether a system is performing as it should, or why a system is performing as it does. This allows internal audits to act as "a mechanism to check that the engineering processes involved in AI system creation and deployment meet declared ethical expectations and standards, such as organizational AI

principles" [34, p. 33]. Internal auditors can also ensure that developers and users of AI systems are acting in accordance with their corporate social responsibility [43]. It has also been recommended that so-called "AGI labs" (typically private organizations purportedly working on artificial general intelligence) must have an internal audit function [44].

- **External/independent auditors**: These are unaffiliated with the company or organization whose AI systems they are auditing and do not have any contractual or financial obligations to it. Given that the auditing is often conducted in the public interest, auditors such as academics or journalists enable an ecosystem of external scrutiny [5], [45], [46]. They are interested in assessing whether the AI system aligns with certain ethical, social, or technical norms and principles.

- **Second-party auditors**: These are typically contracted by the developer of an AI system to audit that system. Such auditors may be granted a degree of access that external auditors may not have, although with mutual agreements on the conditions of disclosure; as such, they may also be described as working collaboratively [42].

- **Communities and users**: AI audits may be conducted directly by individual end-users and by the communities these users belong to. Users may conduct these audits with technical tools developed by experts, through their everyday experiences, or in collaboration with expert auditors [10], [47], [48]. Some auditors have hired Amazon Mechanical Turk workers to stand in for a community of users [49]. Community- and user- driven audits typically take a different view of performance than one afforded by, say, fairness metrics. They assess, often through qualitative means, whether an AI system's

performance meets or impacts the needs of the community conducting the audit [50], [51]. The logistical difficulties of conducting such audits and design challenges affecting their rigor and replicability have meant that few of these audits have been conducted [10], [47]. Existing work has focused on making these audits more effective through sustained collaboration, valuing the expertise of users as auditors, and surfacing users' collective relationships with AI systems and other stakeholders [52], [53]. When community- or user-auditors are motivated to engage adversarially with AI systems, such as by red-teaming LLMs to test their limits and vulnerabilities, they require various legal and technical protections such as safe harbor and freedom from legal action for data scraping [5], [54], [55].

- **Domain experts**: They are specifically described in the literature as suitable for auditing AI systems used in sensitive environments, such as medical algorithms, or in clinical settings [14], [56]. It is suggested that clinical expertise can help derive insights from the operation of "black box" AI systems, especially when these systems fail to align with experts' recommendations and insights.

- **Human-system collaborators**: People might interact with the AI systems—LLMs in particular—being audited and/or another complementary system to identify discrepancies in the audited system's output. What sets these types of audits apart from community or expert audits is that they treat LLMs as collaborators in the audit rather than solely as its target [57]. Sometimes described as co-audits, these help humans detect "mistakes" made by complex AI systems so that they can better align with user intent. Auditors may use human-in-the-loop feedback or "prompt engineering" designed to identify variations or inconsistencies in system output [58], [59], [60].

What and when

The object to be audited often determines the audit's parameters, methodologies, and disclosure mechanisms. These objects include:

- **Sociotechnical system**: AI systems that are subject to audits are typically described as sociotechnical to acknowledge that these technical systems are designed, developed, deployed, used, and governed in social contexts [9], [10], [38], [42], [61], [62], [63]. Audits of sociotechnical AI systems may address (at least one of) these aspects of the system:

    - **Algorithm/model**: Audits examine the type and features of the machine-learning algorithm or model used in the AI system to evaluate its behavior, including by scrutinizing its source code when possible [41], [42], [64], [65]. Such code or model audits are typically used to evaluate whether the design or features of machine-learning models lead to undesirable outcomes, such as biased outputs [20], [62]. Carrying out model audits can be challenging without access to the source code or model features such as weights. This complexity has increased with LLMs, whose input to output path is "non-linear and high-dimensional" [66, p. 1]. Given these challenges, some model audits seek to identify inconsistencies or discrepancies in system outputs rather than in model features. For example, they can assess how outcomes change with tweaks to model features, or how the same input produces different outcomes across different models with similar goals [58], [59], [67]. These audits may also take as their auditing objective the comparison of the observable behavior of a system with formalized criteria [66, p.1]. The release of open-source

models such as OLMo, LLaMa, and Mixtral 8x7B may make model audits or code audits more feasible in the future.

- **Data**: Access to datasets used to train, fine-tune, and/or test the AI system's model can help auditors assess these systems' impacts. As such, some audits of the data used in AI systems examine data flows, provenance, and/or lineage and related dimensions of data quality [69], [70], [71]. These audits also assess the representativeness, diversity, and heterogeneity of data. They evaluate the collection, storage, and use of data for and in AI systems for their adherence to principles of privacy, consent, and sovereignty, such as in the case of Indigenous data rights. More recently, audits have also been used to evaluate the utility and potential for harm of synthetic data: data generated by an AI system that is used for data-analytical applications in real-world contexts [72]. Such audits typically assess synthetic datasets, as well as the models used to generate that data, for their adherence to claims that they protect privacy [73], [74].

- **Interaction with social systems**: Audits might examine AI systems' relationships with the social domains where they are designed, developed, deployed, and used. These relationships might concern private companies, state institutions, marginalized communities, and the general public. These audits assess what impact AI systems have on social processes and systems, their relation to sociocultural biases and sociotechnical harms, their alignment with certain norms and principles or ethical commitments [8], [60], [75] - [79].

- **Governance structures and compliance processes**: Organizations and societies can deploy audits as part of their broader AI management and oversight policies. Falco et al. insist that AI audits "embody three 'AAA' governance principles of prospective risk Assessments, operation Audit trails and system Adherence to jurisdictional requirements" [4, p. 566] Some audits have also been proposed to assess AI systems' compliance with existing legal and regulatory regimes such as the U.K. ICO's AI auditing guidelines, the EU's AIA and audit regimes such as NYC's Local Law 144 [17], [80], [81] . Such audits are usually performed by internal or second-party auditors.

- **Underlying assumptions**: The theoretical foundations, assumptions embedded within datasets, and/or design logics of AI systems are sometimes examined to account for discriminatory or unintended outcomes. Such audits aim to predict the types of impacts that may occur when deploying AI systems in, say, predictive policing or recruitment tools [43], [82].

AI audits vary by their frequency, duration, and by which stage of the AI life cycle they investigate. AI audits can happen as:

- **Snapshot audits**: Most audits are conducted at a specific point in time and test a particular version of an AI system [62]. They are typically conducted either prior to deployment, as in the case of internal audits, or once an AI system has been deployed in real-world conditions, as with external audits.

- **Longitudinal or repeated audits**: Audits may also be conducted over a longer duration than snapshot audit or at regular intervals. Given their logistical difficulty and resource-intensiveness, such audits are scarce [5].

- **Continuous audits**: Another recent proposition in the literature is the continuous monitoring of an algorithmic system according to certain organizationally established criteria and in compliance with regulation [83]. For example, the EU AI Act calls for continuous audits through post-market monitoring [17].

Why

Audits are primarily recommended or conducted for the following reasons:

- **Performance evaluation**: Audits typically ascertain if an AI system is performing according to or diverging from technical and/or statistical baselines[2] as well as a broader set of sociocultural concerns [6], [41], [84]. Performance criteria for an AI audit may account for:

    - **Bias and unfairness**: Audits can assess whether an individual or a group is being treated unfairly, being discriminated against, or is otherwise subject to some form of bias [6], [35], [85]. These biases in system performance can impact processes such as hiring and therefore are a frequent concern of AI audits [75], [86].

    - **Ethics**: Audits may assess an AI system "for consistency with relevant [ethical] principles or norms" [8, p. 324]. These audits, which can overlap with bias and unfairness audits, vary in the ethical principles, such as justice or responsibility, they commit to and how they account for those commitments [87].

    - **Transparency**: AI audits may check whether components of AI systems and/or their governance processes have been documented and made

accessible to internal and external stakeholders [88]. For example, foundation models may be assessed against certain codified indicators of transparency such as the disclosure of data and compute resources and of evaluations of the system's risks and potential harms [89].

- **Interpretability and explainability**: Audits can assess whether an AI system is explainable, that is, whether it can "explain its capabilities and understandings; explain what it has done, what it is doing now, and what will happen next; and disclose the salient information that it is acting on" for the particular task and to user of the system [90, p. 1]. Audits can further assess whether AI systems are inherently interpretable. These audits examine the operations and outputs of AI systems to check if they are comprehensible and useful to people without an additional layer of explanation [91].

- **Robustness and safety**: Audits may help organizations to evaluate whether an AI system is "resilient to attack and adversarial action" that might affect the system's operations and/or harm its users [92, p. 325]. Relatedly, such audits may assess whether an AI system poses public safety risks (such as through their use in healthcare and transportation) [4].

- **Harms discovery**: Audits can assess what possible and actual harms are introduced in social systems or exacerbated by the development and deployment of AI systems [40], [84]. Such harms, given the sociotechnical nature of AI systems might include [38], [51], [78], [93]:
  - stereotypical or harmful language,
  - impeding access to public goods and services,

  - violations of human rights
  - environmental harms that impact local communities

- **Compliance**: Audits may assess AI systems' compliance with legal frameworks and regulations, such as the EU's AIA or data protection guidelines in the U.K., that could impact the development and deployment of AI systems.

- **Privacy**: Audits may assess whether developers' practices of collecting, storing, and/or using data for use in their AI systems protects user privacy [34].

How

To conduct an AI audit the auditor must choose which methodologies, documentation, and tools—often referred to collectively as an audit instrument—will be used to evaluate a particular AI system against the audit criteria discussed above. Typically, an audit assesses the design and operations of a system through qualitative, quantitative (typically statistical), or computational means. The methods and tools used for auditing AI systems may include:

- **Input variation**: In its simplest form, an audit varies the inputs to an AI system to measure and record variations in its outputs [5]. This type of audit corresponds to social–scientific experiments to assess bias, such as those in which an auditor sends an organization CVs with 'white-sounding' or 'Black-sounding' names to investigate differences in hiring decisions. Input variation methods differ in how the audit instrument's inputs (or "prompts" in the literature on LLM audits) are developed and entered into the system, how outcome data are collected, measured, and reported, as well as their complexity. These methods depend on:

- **Automated data generation or collection**: When a system's impacts may be too widespread to investigate qualitatively, auditors assess a system's operations by generating data. For example, sock-puppet methods target particular use cases of an AI system by, "us[ing] computer programs to impersonate users, likely by creating false user accounts or programmatically-constructed traffic" [10, p. 13]. Auditors may also automatically "scrape" platforms supported by AI systems through repeated querying. Based on the scope of the data being collected, these methods range from controlled, which are targeted and limited to specific use-cases of an AI system, to ecological, which involves gathering data on how AI systems interact with their sociotechnical context [5]. Such methods, while often challenged on legal grounds, have been crucial to evaluating AI systems in the absence of publicly available data.

- **Benchmarking/metrics**: These are used to compare the performance of an AI system across inputs and for a certain task [41]. Some build on existing benchmark datasets to audit AI systems, such as WinoBias used in LLMs to assess gender bias [77], [94]. Others are custom-built datasets, for example Raji and Buolamwini's Pilot Parliaments Benchmark, which they developed to audit various Facial Recognition Technologies (FRTs) for the inaccurate recognition of individuals from different demographic groups [6]. Audits may also rely on heterogeneous statistical definitions of "fairness," such as group parity or disparate impact [95], [96]. Audit metrics are also used to test companies' adherence to legally mandated benchmarks for nondiscrimination [21], [35], [42].

- **Automated comparison**: The size of the training datasets and variability of outputs may make large generative models too complex for qualitative and/or statistical comparisons of their outputs. To identify and evaluate output variations, audits of such AI systems therefore employ computational tools such as optimization algorithms or supplementary machine-learning models [59], [66], [67], [68]. When used in conjunction with the system being audited, this method can identify, for example, LLMs' biased predictions, toxic rhetoric or "inauthentic" outputs generated by Generative Adversarial Networks (GANs) [66], [73]. Automated audit instruments require significant technical expertise and resources to build and use.

- **Audit artifacts**: Records, logs, audit trails, and similar tangible documentation may be used to check whether an organization using or developing an AI system adheres to a certain set of guidelines, ethical principles, or organizational policy in a traceable manner. For AI systems, audit artifacts may include model cards that are designed to explain various features of a machine-learning model and/or data sheets that describe the training data used for AI systems, how they were tested for bias, and so on [6], [97], [98].

- **Guiding questions**: Some audits evaluate AI systems through a set of questions, e.g. "under what conditions the system makes mistakes or inaccurate predictions?" and corresponding checklists directed to the system and its stakeholders [14], [79].

- **Toolkits**: Auditors may use standardized toolkits specifically designed for audits. Some, such as AI Fairness360 and Aequitas are software packages that

audit the performance of AI systems according to provided fairness criteria [95], [96]. Others provide procedures for audits to meet regulatory requirements, such as capAI [99]. Still others, such as the Algorithmic Equity Toolkit, use participatory design to enable community stakeholders to audit these systems [50].

- **Agent-based simulations**: These simulate the social context of an AI system and the actors in it to measure potentially negative impacts, such as income disparity [100]. These recently developed methods are not, however, widely taken up in the literature.

These methods, and the decision to deploy them, vary according to the level of access an auditor has to the system [34]. These can range from black-box access, where auditors have no access to the internal operations of an AI system, to white-box access, which provides auditors complete access to the system.

What next

Companies and organizations, regulators, impacted communities, and auditors themselves must often consider what can or should be done after an audit. Some of the recommended post-audit processes are listed below:

- **Reporting**: Internal auditors typically share their results with the organization tasked with managing an AI system. Second-party or external auditors usually publish their findings in scientific journals and/or news publications. The public disclosure of audit results is important for AI accountability and trustworthiness [5], [101].
- **User notification**: Costanza-Chock et al. recommend that those directly or indirectly affected by AI systems be informed of any impacts [101]. Notifications

should be designed to let affected persons opt out of the system without affecting their essential user experience (without, for example, impeding access to vital public services).

- **Mitigation strategies**: Developers or deployers of AI systems may build mechanisms to mitigate or protect against identified adverse impacts [34]. For example, they may develop forms of interpretability or explainability for the system, debias algorithms, retrain the model using debiased data, or end the system's use.

- **Certification**: Based on the results of an AI audit and the audit target's responses to them, regulator- or industry-backed certifications may be granted. Assurance mechanisms may demonstrate to various stakeholders the AI system's adherence to normative principles, such as fairness or responsibility, and/or its compliance with regulations or standards [34], [43].

Table 1 summarizes how organizational, institutional, and legal frameworks and standards for AI governance and auditing answer these guiding questions.

Table 1. How Various Audit Regimes Answer Proposed Guiding Questions.

Audit regime	Instituting entity	Who	What (and when)	Why	How	What next
AI Auditing Framework	Institute of Internal Auditors	Internal auditors	Sociotechnical system; unclear when	Governance, performance	Audit artifacts, guiding questions	Undefined
AI RMF 1.0	National Institute of Standards & Technology, U.S.A	Undefined	Sociotechnical system; Repeated	Governance (risk, trustworthiness), performance, harm	Undefined	Undefined
Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities	Government Accountability Office, U.S.A	Internal and external auditors	Sociotechnical system, governance structures and processes; Repeated	Governance (accountability), performance	Audit artifacts, guiding questions	Undefined
ML4H Audit Framework	ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H)	Internal auditors and domain experts	Sociotechnical system, governance structures and processes; Snapshot	Performance (bias, robustness, interpretability)	Input variation (metrics), audit artifacts, guiding questions	Reporting
ISO/IEC 42001:2023	ISO	Internal auditors	Sociotechnical system, governance structures and processes; Repeated	Governance (responsibility), compliance / conformity	Audit artifacts, guiding questions	Reporting
Model AI Governance Framework	Infocomm Media Development Authority, Singapore	Internal or external auditors	Sociotechnical system; Snapshot or repeated	Governance (risk), performance	Audit artifacts	Reporting
Audit framework for algorithms	Netherlands Court for Audit	External auditors	Sociotechnical system; Snapshot	Performance (bias, ethics), compliance	Guiding questions	Reporting

Digital Services Act and Online Safety Act	European Union	Internal, second-party, or external auditors	Algorithm; Repeated	Performance, compliance	Undefined	Reporting
AI Act	European Union	Internal, second-party, or external auditors	Sociotechnical system; Repeated	Governance (risk), performance (bias), harms discovery, compliance	Undefined	Reporting, certification
Code of Practice for Generative AI Systems Developers, Deployers, and Operators	Innovation, Science, and Economic Development, Canada	Internal and external auditors	Undefined	Compliance	Undefined	Undefined
Algorithmic Impact Assessment	Treasury Board, Canada	Internal auditors	Sociotechnical system; Snapshot	Governance (risk), performance (bias), compliance	Guiding questions	Reporting
Local Law 144	New York City	External auditors	Sociotechnical system; Repeated	Performance (bias), compliance	Input variation (metrics)	Reporting
AI Audits	Information Commissioner's Office, United Kingdom	External auditors	Data; Snapshot	Compliance	Audit artifacts, guiding questions	Reporting

Source: IPIE analysis based on this literature review.

SECTION 3. DISCUSSION AND CONCLUSION

As the previous sections make clear, audits differ by who conducts them, what is audited and when, why the audit is conducted, how it is conducted, and what actions are taken afterwards. The differences determine what an audit can tell us about an AI system's governance and performance.

The type of auditor, internal or external to the organization, significantly impacts what aspects of an AI system can be audited. This includes the scope of access to audit evidence, the methodology employed, the auditor's independence from the system being audited, and to whom they report. Auditors leverage a diverse set of methodologies, tools, and procedures to evaluate an AI system against criteria carefully chosen based on the audit's specific objectives.

In terms of objects, audits assess the algorithms/models, the data, and the impact of AI system's decisions, predictions, or other outcomes on stakeholders. Their scope varies depending on which stage(s) of the AI life cycle—from design to operations and monitoring—they examine and the system's version, both of which also impact their duration—snapshot, longitudinal, or continuous.

Broadly, audits seek to assess if an AI system and its governance conform with regulation, international standards, and/or if a system performs according to, or deviates from, an acceptable baseline. These aims shape evaluation criteria; for example, an audit to assess a hiring tool's potential bias may test its performance against legal measures of non-discrimination. These aims also determine an audit's parameters. For example, checking if an AI system makes racially discriminatory predictions does not require access to the system's machine-learning model(s) but making strong claims about why those outcomes arise might require that access.

All the above factors impact how an audit can be conducted, including the choice of audit instrument and methodology. For example, because deployers of AI systems value their regulatory compliance, internal or second-party auditors are granted significant access to the system. This allows auditors to rely on standard methods and tools, such as guided interviews with product managers and audit artifacts maintained by the company. In contrast, to assess a system's societal impacts and potential harms, independent auditors may use methods such as input variation, which do not require such cooperation.

After the audit has been completed, its outcomes are typically shared with those managing the AI system and/or those stakeholders who commissioned it. While resulting certification and other assurance mechanisms may demonstrate the trustworthiness of AI systems, the difficulty of verifying internal audit results remains a concern for external stakeholders.

Based on this review, we identify three key takeaways about the landscape of AI auditing:

1. To accurately assess the potential risks, and impacts of AI systems, we need a trustworthy audit ecosystem with complementary approaches from internal, external, and community auditors.

   Effective internal audits assess organization structures and processes for their compliance with AI governance policies and regulation, as well as monitor them for concerns, such as algorithmic bias and training data quality. Audit firms, who have developed specialized methodologies and tools, can perform second-party audits to help organizations.

   Limiting the audit ecosystem to internal and second-party audits, however, can undermine public trust in AI systems and developer accountability. While

regulatory features such as registries of audit results and auditor accreditation can help ensure audit quality, the limited verifiability of internal and second-party audits means that there remain concerns about auditors' privileged access to AI systems; auditors' independence from the companies whose systems they audit; the risk of "audit-washing" (in ways analogous to ethics-washing) AI systems that still pose significant problems to societal wellbeing [92], [102], [103]. For example, since the results of the bias audits mandated by New York City's Local Law 144 have not been widely disclosed, it is extremely difficult to verify whether AI systems violate the conditions of that law and if companies have taken any bias mitigation measures; a form of non-verifiable compliance that Wright et al. term "null compliance" [104].

To allay the concerns outlined above, other types of audits, such as independent assessments conducted by academic researchers, civil society actors, and impacted communities, can be used to verify the claims made by the developers and deployers of AI systems, as well as publicly disclose harms they've discovered. These audits often have different criteria, aims, and methodologies than those of internal or second-party audits, which means they can reveal previously unknown impacts and outcomes of AI systems [16]. These audits can also employ adversarial methodologies such as red-teaming that require legal safeguards [55]. Independent auditors, however, do not have unimpeded access to the system, which can limit audit scope. Also given that these audits often target specific parts of an AI system, they are not intended to be comprehensive assessments of the system.

Evidently, internal, second-party and external/independent audits each have their strengths and limitations. A diverse audit ecosystem can therefore help harness the opportunities of AI systems while limiting their risks.

2. Auditors need better access to data and audit artifacts from the developers and deployers of AI systems. Comprehensive auditability requires documentation and disclosure of an AI system's model and data components, associated risks and impacts, and easily understandable explanations of its outcomes. While some audits have previously been conducted with limited system access, AI systems' global impacts are difficult to assess without documentation of system features [88]. Internal auditors typically have greater access to an AI system than external auditors, however, auditability remains a significant concern to both.

3. Given that the development and use of AI systems impacts communities across the world, audit regimes must account for their global effects. Most existing audits have been conducted in North America, Europe, and other regions of the 'global north', with their results typically published in English and focused on effects within these regions [105]. The impacts of AI systems, however, include shifts in social and environmental conditions beyond the immediate development or application contexts of a system. While several countries have demonstrated interest in audits to assess AI systems' impacts (see Appendix B), effective AI governance requires more comprehensive audits that address these widespread shifts. AI systems affect labor through the automation of work, knowledge extraction, and the precarious and stressful working conditions of data annotators [106], [107], [108]. The training and use of these systems consumes massive quantities of water and energy [109]. These issues become a matter of greater global concern given how far AI supply chains stretch. Audit regimes, then, need to reflect the actual scope of AI systems.

As indicated in the introduction, this review will inform the development of the criteria, evaluation mechanisms, evidentiary documentation, and processes crucial

to global standards for AI audits. Standards can build trust in the quality and rigor of different audits as well as make these audits' results comparable. Additionally, they can help balance power dynamics between those invested in AI systems (often in the global north) and auditors by laying out comprehensively the terms to which AI developers and deployers should be held accountable. To be effective, however, global standards must address issues related to the trustworthiness and scope of AI audits. Standardization should not erase the important differences between audits conducted by different auditors with specific aims, criteria, and methodologies. Furthermore, standardization should not overlook the differences between the impacts of AI systems on local communities and on a global scale. Having outlined these concerns, we describe in a separate report the protocols needed to inform the development of global standards for AI audits.

ENDNOTES

[1] The use of AI systems here reflects their sociotechnicality: the social, systemic, and technical components that ground the collection and use of data, the development and deployment of algorithms as AI, and their application context. However, it is important to note that current AI discourse and audits that we review leans towards data-driven AI system (such as Large Language Models, or LLMs) rather than, say, rule-based models.

[2] Performance evaluation metrics used in AI audits may include, but are not limited to, metrics such as accuracy, precision, recall, and regression metrics that are used to evaluate model performance.

REFERENCES

[1] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, "Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, Art. no. CSCW1, Apr. 2021, doi: 10.1145/3449081.

[2] 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

[3] *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, National Institute of Standards and Technology, USA, Jan. 26, 2023.

[4] 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.

[5] 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.

[6] 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.

[7] S. Brown, J. Davidovic, and A. Hasan, "The algorithm audit: Scoring the algorithms that score us," *Big Data & Society*, vol. 8, no. 1, Art. no. 1, Jan. 2021, doi: 10.1177/2053951720983865.

[8] J. Mökander and L. Floridi, "Ethics-Based Auditing to Develop Trustworthy AI," *Minds & Machines*, vol. 31, no. 2, Art. no. 2, Jun. 2021, doi: 10.1007/s11023-021-09557-8.

[9] B. Vecchione, K. Levy, and S. Barocas, "Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, -- NY USA: ACM, Oct. 2021, pp. 1–9. doi: 10.1145/3465416.3483294.

[10] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. "Auditing algorithms: Research methods for detecting discrimination on internet platforms". In Data and Discrimination: Converting Critical Concerns into Productive Inquiry, a preconference at the 64th Annual Meeting of the International Communication Association, 2014, pp. 4349–4357.

[11] M. Power, *The audit society: rituals of verification*, Repr. Oxford: Oxford University Press, 2013.

[12] ISACA, *Auditing Artificial Intelligence,* Schaumburg, IL: ISACA, 2018.

[13] Institute of Internal Auditors, *The IIA's Artificial Intelligence Auditing Framework*, 2023.

[14] 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

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

[16] P. Terzis, M. Veale, and N. Gaumann, "Law and the emerging political economy of algorithmic audits." In 2024 ACM Conference on Fairness, Accountability, and Transparency (FACCT '24), June 03–06, 2024, Rio de Janeiro, Brazil, 2024. ACM. doi:10.1145/3630106.3658970, Available at SSRN: https://ssrn.com/abstract=4794565

[17] 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.

[18] M. Sheehan and S. Du, "What China's algorithm registry reveals about AI governance", *Carnegie Endowment for International Peace,* 2022. [Online]. Available: https://carnegieendowment.org/2022/12/09/what-china-s-algorithm-registry-reveals-about-ai-governance-pub-88606.

[19] Innovation, Science and Economic Development Canada, *Canadian guardrails for generative AI — Code of practice.* Government of Canada, 2023. [Online]. Available: https://ised-isde.canada.ca/site/ised/en/consultation-development-canadian-code-practice-generative-artificial-intelligence-systems/canadian-guardrails-generative-ai-code-practice

[20] 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.

[21] 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

[22] 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.

[23] I. Saif and B. Ammanath, ""Trustworthy AI" is a framework to help manage unique risk", *MIT Technology Review,* 2020. [Online]. Available: https://www.technologyreview.com/2020/03/25/950291/trustworthy-ai-is-a-framework-to-help-manage-unique-risk/.

[24] KPMG, "The KPMG trusted AI approach", 2023. [Online]. Available: https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2023/12/kpmg-trusted-ai-approach.pdf.

[25] PwC, "Responsible AI – Maturing from theory to practice," 2020. [Online]. Available: https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai/pwc-responsible-ai-maturing-from-theory-to-practice.pdf

[26] EY, "EY's commitment to developing and using AI ethically and responsibly," 2023. [Online]. Available: https://www.ey.com/en_gl/insights/ai/principles-for-ethical-and-responsible-ai

[27] K. Buehler, R. Dooley, L. Grennan, and A. Singla, "Getting to know—And manage—Your biggest AI risks", *McKinsey Analytics,* 2021. [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/getting-to-know-and-manage-your-biggest-ai-risks#/

[28] BCG, "Responsible AI," BCG Global. [Online]. Available: https://www.bcg.com/capabilities/artificial-intelligence/responsible-ai.

[29] Meta, "Facebook's five pillars of responsible AI", 2021. [Online]. Available: https://ai.meta.com/blog/facebooks-five-pillars-of-responsible-ai/

[30] Google, "Introduction to responsible AI," 2023. [Online]. Available: https://developers.google.com/machine-learning/resources/intro-responsible-ai.

[31] Amazon, "Our commitment to the responsible use of AI". 2023. [Online]. Available: https://www.aboutamazon.com/news/company-news/amazon-responsible-ai.

[32] OpenAI, "Our approach to AI safety," 2023. [Online]. Available: https://openai.com/index/our-approach-to-ai-safety/

[33] Meta "FAIR progress and learnings across socially responsible AI research," *AI at Meta,* 2023. [Online]. Available: https://ai.meta.com/blog/fair-progress-and-learnings-across-socially-responsible-ai-research/.

[34] 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.

[35] C. O'Neil, H. Sargeant, and J. Appel, "Explainable Fairness in Regulatory Algorithmic Auditing," Oct. 10, 2023, *Rochester, NY*: 4598305. Accessed: Nov. 16, 2023. [Online]. Available: https://papers.ssrn.com/abstract=4598305

[36] G. Galdon Clavell, M. Martín Zamorano, C. Castillo, O. Smith, and A. Matic, "Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization," in *Proceedings of*

the AAAI/ACM Conference on AI, Ethics, and Society, New York NY USA: ACM, Feb. 2020, pp. 265–271. doi: 10.1145/3375627.3375852.

[37] Infocomm Media Development Authority, *Model Artificial Intelligence governance framework (second edition),* 2020. [Online]. Available: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

[38] 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.

[39] T. Markert, F. Langer, and V. Danos, "GAFAI: Proposal of a Generalized Audit Framework for AI," In Lecture Notes in Informatics (LNI) from INFORMATIK 2022, 2022, pp. 1247-1256. doi: 10.18420/INF2022_107.

[40] J. Bandy, "Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, pp. 1–34, Apr. 2021, doi: 10.1145/3449148.

[41] 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.

[42] 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.

[43] I. Ajunwa, "An Auditing Imperative for Automated Hiring," Mar. 15, 2019, *Rochester, NY*: 3437631. doi: 10.2139/ssrn.3437631.

[44] J. Schuett, "AGI labs need an internal audit function," May 26, 2023, *arXiv*: arXiv:2305.17038. [Online]. Available: http://arxiv.org/abs/2305.17038

[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] M. Broussard, "How to Investigate an Algorithm," *Issues*, vol. 39, no. 4, pp. 85–89, Jul. 2023, doi: 10.58875/OAKE4546.

[47] 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.

[48] H. Shen, A. DeVos, M. Eslami, and K. Holstein, "Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–29, Oct. 2021, doi: 10.1145/3479577.

[49] 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.

[50] 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.

[51] 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.

[52] W. H. Deng, B. Guo, A. Devrio, H. Shen, M. Eslami, and K. Holstein, "Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–18. doi: 10.1145/3544548.3581026.

[53] A. DeVos, A. Dhabalia, H. Shen, K. Holstein, and M. Eslami, "Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior," in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–19. doi: 10.1145/3491102.3517441.

[54] N. Inie, J. Stray, and L. Derczynski, "Summon a Demon and Bind it: A Grounded Theory of LLM Red Teaming in the Wild," 2023, *arXiv*. doi: 10.48550/ARXIV.2311.06237.

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

[56] 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.

[57] A. D. Gordon *et al.*, "Co-audit: tools to help humans double-check AI-generated content," Oct. 02, 2023, *arXiv*: arXiv:2310.01297. [Online]. Available: http://arxiv.org/abs/2310.01297

[58] 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

[59] M. Amirizaniani *et al.*, "Developing a Framework for Auditing Large Language Models Using Human-in-the-Loop," Feb. 16, 2024, *arXiv*: arXiv:2402.09346. [Online]. Available: http://arxiv.org/abs/2402.09346

[60] J. H. Rystrøm, "Apolitical Intelligence? Auditing Delphi's responses on controversial political issues in the US," Jun. 22, 2023, *arXiv*: arXiv:2306.13000. Accessed: May 07, 2024. [Online]. Available: http://arxiv.org/abs/2306.13000

[61] 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.

[62] 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.

[63] E. Radiya-Dixit and G. Neff, "A Sociotechnical Audit: Assessing Police Use of Facial Recognition," in *2023 ACM Conference on Fairness, Accountability, and Transparency*, Chicago IL USA: ACM, Jun. 2023, pp. 1334–1346. doi: 10.1145/3593013.3594084.

[64] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness," in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, Jul. 2018, pp. 2564–2572. Accessed: Jan. 04, 2024. [Online]. Available: https://proceedings.mlr.press/v80/kearns18a.html

[65] 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.

[66] E. Jones, A. Dragan, A. Raghunathan, and J. Steinhardt, "Automatically Auditing Large Language Models via Discrete Optimization," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 2023, pp. 15307–15329. Accessed: May 07, 2024. [Online]. Available: https://proceedings.mlr.press/v202/jones23a.html

[67] H. Bharadhwaj, D.-A. Huang, C. Xiao, A. Anandkumar, and A. Garg, "Auditing AI models for Verified Deployment under Semantic Specifications," Nov. 01, 2021, *arXiv*: arXiv:2109.12456. Accessed: May 07, 2024. [Online]. Available: http://arxiv.org/abs/2109.12456

[68] M. L. Olson, S. Liu, R. Anirudh, J. J. Thiagarajan, P.-T. Bremer, and W.-K. Wong, "Cross-GAN Auditing: Unsupervised Identification of Attribute Level Similarities and Differences Between Pretrained Generative Models," in *2023 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 7981–7990. doi: 10.1109/CVPR52729.2023.00771.

[69] 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. [Online]. Available: http://arxiv.org/abs/2310.16787

[70] V. Gudivada, A. Apon, and J. Ding. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations". In: International Journal on Advances in Software 10.1 (2017), pp. 1 - 20.

[71] T. Pasquier, J. Singh, J. Powles, D. Eyers, M. Seltzer, and J. Bacon, "Data provenance to audit compliance with privacy policy in the Internet of Things," *Pers Ubiquit Comput*, vol. 22, no. 2, pp. 333–344, Apr. 2018, doi: 10.1007/s00779-017-1067-4.

[72] J. Jordon *et al.*, "Synthetic Data -- what, why and how?," 2022, *arXiv*. doi: 10.48550/ARXIV.2205.03257.

[73] A. Alaa, B. V. Breugel, E. S. Saveliev, and M. van der. Schaar, "How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models," In *Proceedings of the 39th International Conference on Machine Learning*, PMLR, 2023, pp. 290–306. https://proceedings.mlr.press/v162/alaa22a.html

[74] B. Belgodere *et al.*, "Auditing and Generating Synthetic Data with Controllable Trust Trade-offs," 2023, *arXiv*. doi: 10.48550/ARXIV.2304.10819.

[75] L. Armstrong, A. Liu, A., S. MacNeil and D. Metaxa, "The Silicone Ceiling: Auditing GPT's Race and Gender Biases in Hiring," 2024, *arXiv preprint arXiv:2405.04412*.

[76] J. D. Gaebler, S. Goel, A. Huq, and P. Tambe, "Auditing the Use of Language Models to Guide Hiring Decisions," Apr. 03, 2024, *arXiv*: arXiv:2404.03086. Accessed: May 07, 2024. [Online]. Available: http://arxiv.org/abs/2404.03086

[77] H. Kotek, R. Dockum, and D. Sun, "Gender bias and stereotypes in Large Language Models," in *Proceedings of The ACM Collective Intelligence Conference*, Delft Netherlands: ACM, Nov. 2023, pp. 12–24. doi: 10.1145/3582269.3615599.

[78] 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*, Montr\'{e}al QC Canada: ACM, Aug. 2023, pp. 723–741. doi: 10.1145/3600211.3604673.

[79] A., Kennedy, D. Coates, and K. Lindquist, "Auditing Government AI: How to assess ethical vulnerability in machine learning," In *Workshop on Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020),* 2020. https://ai-broader-impacts-workshop.github.io/papers/audtinggovtai.pdf

[80] E. Kazim, D. M. T. Denny, and A. Koshiyama, "AI auditing and impact assessment: according to the UK information commissioner's office," *AI Ethics*, vol. 1, no. 3, Art. no. 3, Aug. 2021, doi: 10.1007/s43681-021-00039-2.

[81] 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.

[82] P. Ugwudike, "AI audits for assessing design logics and building ethical systems: the case of predictive policing algorithms," *AI Ethics*, vol. 2, no. 1, pp. 199–208, Feb. 2022, doi: 10.1007/s43681-021-00117-5.

[83] M. Minkkinen, J. Laine, and M. Mäntymäki, "Continuous Auditing of Artificial Intelligence: a Conceptualization and Assessment of Tools and Frameworks," *DISO*, vol. 1, no. 3, p. 21, Dec. 2022, doi: 10.1007/s44206-022-00022-2.

[84] 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

[85] J. Asplund, M. Eslami, H. Sundaram, C. Sandvig, and K. Karahalios. "Auditing race and gender discrimination in online housing markets," In *Proceedings of the international AAAI conference on web and social media,* 2020, pp. 24-35.

[86] L. Lippens, "Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, p. 100054, Jan. 2024, doi: 10.1016/j.chbah.2024.100054.

[87] J. Laine, M. Minkkinen, and M. Mäntymäki, "Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders," *Information & Management*, vol. 61, no. 5, Jul. 2024, doi: 10.1016/j.im.2024.103969.

[88] 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

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

[90] 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.

[91] 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.

[92] E. P. Goodman and J. Trehu, "Algorithmic Auditing: Chasing AI Accountability,” *Santa Clara High Tech. LJ*, vol. 39, no. 1., 2023, pp. 289-338

[93] J. Haider, M. Rödl, and S. Joosse, “Algorithmically embodied emissions: the environmental harm of everyday life information in digital culture,” *IR*, vol. 27, 2022, doi: 10.47989/colis2224.

[94] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 15–20. doi: 10.18653/v1/N18-2003.

[95] P. Saleiro *et al.*, “Aequitas: A Bias and Fairness Audit Toolkit,” Apr. 29, 2019, *arXiv*: arXiv:1811.05577. Accessed: Jan. 18, 2024. [Online]. Available: http://arxiv.org/abs/1811.05577

[96] R. K. E. Bellamy *et al.*, “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias,” Oct. 03, 2018, *arXiv*: arXiv:1810.01943. Accessed: Jan. 31, 2024. [Online]. Available: http://arxiv.org/abs/1810.01943

[97] M. Mitchell *et al.*, “Model Cards for Model Reporting,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta GA USA: ACM, Jan. 2019, pp. 220–229. doi: 10.1145/3287560.3287596.

[98] T. Gebru *et al.*, “Datasheets for Datasets,” Dec. 01, 2021, *arXiv*: arXiv:1803.09010. Accessed: Jan. 23, 2024. [Online]. Available: http://arxiv.org/abs/1803.09010

[99] 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.

[100] 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.

[101] 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.

[102] O. Keyes, J. Hutson, and M. Durbin, "A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–11. doi: 10.1145/3290607.3310433.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] M. Pasquinelli and V. Joler, "The Nooscope manifested: AI as instrument of knowledge extractivism," *AI & Soc*, vol. 36, no. 4, pp. 1263–1280, Dec. 2021, doi: 10.1007/s00146-020-01097-6.

[108] M. L. Gray and S. Suri, *Ghost work: how to stop Silicon Valley from building a new global underclass*. Boston: Houghton Mifflin Harcourt, 2019.

[109] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, Mar. 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[110] Ministry of Science, Technology, and Innovations, *Summary of the Brazilian Artificial Intelligence strategy—EBIA,* 2021. [Online]. Available: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ebia-summary_brazilian_4-979_2021.pdf

[111] Ministry of Science, Technology, Knowledge and Innovation, *National policy on Artificial Intelligence,* 2021. [Online]. Available: https://minciencia.gob.cl/uploads/filer_public/bc/38/bc389daf-4514-4306-867c-760ae7686e2c/documento_politica_ia_digital_.pdf

[112] Webster, G., & Laskai, L. (Trans.), "*T*ranslation: Chinese expert group offers "governance principles" for "responsible AI"," *DigiChina,* 2019. [Online]. Available:

https://digichina.stanford.edu/work/translation-chinese-expert-group-offers-governance-principles-for-responsible-ai/

[113] Ministry of Industry and Trade of the Czech Republic, *National Artificial Intelligence strategy of the Czech Republic,* 2019. [Online]. Available: https://www.mpo.gov.cz/assets/en/guidepost/for-the-media/press-releases/2019/5/NAIS_eng_web.pdf

[114] Ministry of Economic Affairs and Communications, *National AI strategy 1.0,* 2021. [Online]. Available: _https://www.kratid.ee/_files/ugd/980182_8d0df96fd41145739dff2595e0ab3e8d.pdf

[115] Villani, C., Bonnet, Y., Berthet, C., Levin, F., Schoenauer, M., Cornut, A. C., & Rondepierre, B, *Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne,"* Conseil national du numérique, 2018.

[116] DKE German Commission for Electrical, Electronic & Information Technologies of DIN and VDE, *German standardization roadmap on Artificial Intelligence, 2nd edition,* 2023. [Online]. Available: https://www.din.de/resource/blob/916798/ed09ae58b60f0d3a498fa90fa5085b7c/nrm-ki-engl-2023-final-web-250-neu-data.pdf

[117] NITI Aayog, *Responsible AI #AIForAll Approach Document for India,* 2021. [Online]. Available: https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf.

[118] Badan Pengkajian dan Penerapan Teknologi, *Indonesia national AI strategy (Stranas KA) 2020-2045,* 2020. [Online]. Available: https://ai-innovation.id/images/gallery/ebook/stranas-ka.pdf

[119] Department of Enterprise, Trade and Employment, *AI – Here for good: A national Artificial Intelligence strategy for Ireland,* 2021. [Online]. Available: https://enterprise.gov.ie/en/publications/publication-files/national-ai-strategy.pdf

[120] Integrated Innovation Strategy Promotion Council (Cabinet Office) & AI Strategy Council, *National AI strategy – 2022,* 2022. [Online]. Available*:* https://www8.cao.go.jp/cstp/ai/aistratagy2022en.pdf

[121] IA2030Mx., *Mexican national AI agenda,* 2018. [Online]. Available: https://wp.oecd.ai/app/uploads/2022/01/Mexico_Agenda_Nacional_Mexicana_de_IA_2030.pdf

[122] Netherlands Court of Audit, *Understanding algorithms,* 2021. [Online] Available: https://english.rekenkamer.nl/binaries/rekenkamer-english/documenten/reports/2021/01/26/understanding-algorithms/Understanding+algorithms+-+2021.pdf

[123] Fundação para a Ciência e Tecnologia, *AI Portugal 2030 – National strategy for AI,* 2019. [Online]. Available: https://wp.oecd.ai/app/uploads/2021/12/Portugal_AI_Portugal_2030_2019-2030.pdf

[124] MINECO, *National Artificial Intelligence strategy,* 2019. [Online]. Available: https://portal.mineco.gob.es/RecursosArticulo/mineco/ministerio/ficheros/National-Strategy-on-AI.pdf

[125] Presidency of the Republic of Turkey's Digital Transformation Office & Ministry of Industry and Technology, *National AI Strategy 2021-2025,* 2021. [Online]. Available: https://cbddo.gov.tr/SharedFolderServer/Genel/File/TRNationalAIStrategy2021-2025.pdf

[126] Agesic, *Artificial intelligence strategy for the digital government, 2019.* [Online]. Available: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/sites/agencia-gobierno-electronico-sociedad-informacion-conocimiento/files/documentos/publicaciones/IA%20Strategy%20-%20english%20version.pdf

[127] OECD.AI, powered by EC/OECD, *database of national AI policies*, 2021. [Online]. Available: https://oecd.ai.

APPENDIX A: A NOTE ON METHODOLOGY

This literature review identifies key academic sources focused on AI/algorithm(ic) audits primarily through keyword searches on Google Scholar, Scopus, and Web of Science. In order to include work from machine learning, computer science, computational linguistics, and allied fields that frequently circulate as pre-prints, we also searched the arXiv database. We used the following search terms: AI audit, algorithm audit, algorithmic audit, LLM audit, generative AI audit. We also identified highly influential journal articles based on citation counts and references in the literature, and expert recommendations. We disregarded work that described the use of AI systems to conduct other forms of audits, such as internal audits and financial accounting. We also did not consider cybersecurity audits, whose primary concerns differ from those of the literature on AI/algorithm audits.

We also searched regulation, legislation, and policy documents related to AI by keyword on Google Search and manually identified those documents of interest to AI audits. We used the above-mentioned search terms, appended with AI responsibility, AI accountability, trustworthy AI, ethical AI, AI safety, AI policy, and AI governance. We also identified related documentation through references in academic work and expert recommendations. To collect frameworks, guidelines, and standards around AI audits, we searched databases of auditor bodies such as the Institute of Internal Auditors and the Information Systems Audit and Control Association (ISACA), standard-setting bodies such as the International Organization for Standardization (ISO) and NIST (National Institute of Standards and Technology), and custom databases developed by entities concerned with AI policy and governance such as the AI Standards Hub (developed by the Alan Turing Institute, the British Standards Institution, and the National Physical Laboratory) and the OECD's AI Policy Observatory. We also searched the databases of the "big 4" audit firms such as

Deloitte and KPMG, and firms focused on AI/algorithm auditing such as Holistic AI and Eticas for relevant literature.

Following our keyword search, we collected 78 articles published in peer-reviewed journals and as preprints, 21 documents from industry associations and standard-setting organizations, and national policy documents and regulation from 20 countries. All texts in our corpus were read in full and manually analyzed for their relevance and contribution to the scholarly, industrial, and governmental perspectives on AI audits.

APPENDIX B: NATIONAL AI POLICIES THAT RECOMMEND AUDITS

Table 2 shows governmental guidelines/frameworks for AI systems that incorporate audits. While many of the audit regimes described and/or recommended here are similar to those suggested in regulations and/or legislation described in the main body of the literature review, they are non-binding and lack enforcement or accountability mechanisms.

Table 2. Policy Documents Related to AI Governance that Incorporate Audits.

Country	Policy title	Issuing body or bodies	Relevance to AI audits
Brazil	Brazilian AI Strategy	Ministry of Science, Technology, and Innovations	Recognizes the importance of making AI systems "more traceable, auditable and accountable" [110, p. 11].
Chile	National Policy on Artificial Intelligence	Ministry of Science, Technology, Knowledge and Innovation	Recognizes that in order to address the ethical challenges of AI, specifically in the domain of gender equality, AI systems must be audited [111].
China	Governance Principles for New Generation AI: Develop Responsible Artificial Intelligence	National New Generation AI Governance Expert Committee	Asserts that auditability must be a goal for safe/secure and controllable AI systems [112].
Czech Republic	National Artificial Intelligence Strategy of the Czech Republic	Ministry of Industry and Trade of the Czech Republic	Proposes the development of a "certified methodology for system audits in co-operation of the public and private sectors" [113, p. 36].
Estonia	National AI Strategy 1.0	Ministry of Economic Affairs and Communications	While the strategy generally touches upon aspects of responsible AI use, it specifically recommends supporting public sector AI use through "data audits" [114, p. 5].
France	Stratégie Nationale pour l'intelligence artificielle (SNIA)	Prime Minister	Notes the importance of audits and AI systems' auditability to the transparency of these systems. Notably, the report on which this strategy is based advocates for "citizen audits" through open access to data [115, p. 45].

Germany	German Standardization Roadmap on Artificial Intelligence	DKE German Commission for Electrical, Electronic & Information Technologies of DIN and VDE	Notes conformity assessments, as laid out in the EU's AI Act, as a means to 'test' and 'verify' the ethical and performance dimensions of AI systems, among others. Auditing, however, is explicitly limited to organizations and does not apply to systems/products [116].
India	Responsible AI #AIForAll Approach Document for India	NITI Aayog	In the "Principles for Responsible Management of AI Systems" that have been developed, audits are noted as a mechanism that should ensure that the design, development, and deployment of AI systems adhere to the stated principles of transparency and accountability. These audits can be done by internal, external, or – notably – public entities, such as research institutions [117].
Indonesia	Indonesia National AI Strategy (Stranas KA) 2020–2045	Badan Pengkajian dan Penerapan Teknologi (Agency for the Assessment and Application of Technology)	Audits are recognized as an aspect of the data and infrastructure ecosystems for AI systems. It is recommended that these audits are conducted by the Agency for the Assessment and Application of Technology in accordance with the Perlindungan Data Pribadi, or Personal Data Protection, regulation until a body dedicated to AI policy is established [118]. Such an organization, aligned with the 'Quadruple-Helix' collaborative ecosystem consisting of government, industry, academia, and community, would be committed to upholding transparent, professional, audited, and responsible AI governance.
Ireland	AI – Here for Good: A National Artificial Intelligence Strategy for Ireland	Department of Enterprise, Trade and Employment	As part of the recommended effort to establish standards and certification for "trustworthy AI", audits are suggested as assurance tools and as part of certification schemes and codes of conduct [119].
Japan	AI Strategy -2022	Integrated Innovation Strategy Promotion Council (Cabinet Office), AI Strategy Council	Recommends audits as a mechanism to ensure AI responsibility [120].

Mexico	Artificial Intelligence in Mexico: A National Agenda	IA2030Mx	Recommends the establishment of ethical standards that AI systems can be audited against [121]. Also suggests incentivizing the establishment of independent auditing organizations.
Netherlands	Audit Framework for Algorithms	Netherlands Court of Audit	The Netherlands Court of Audit has developed an audit framework that it used to investigate nine algorithms used by the Dutch government. This framework considers dimensions of the algorithm's "governance and accountability, model and data, privacy, quality of IT General Controls [such as access and back-up controls], and ethics" [122, p. 23]
Portugal	AI Portugal 2030 – National Strategy for AI	Foundation for Science and Technology (FCT)	Recommends traceability and auditability as means of ensuring privacy, fairness, and other safety and ethical principles [123].
Spain	National Artificial Intelligence Strategy	Ministry of Economy, Trade and Business f.k.a Ministry of Economic Affairs and Digital Transformation (MINECO)	Emphasizes that AI systems must be "transparent and auditable" so that how they work and the decisions they make can be "audited, evaluated and explained" [124, p. 66- 67].
Türkiye	National Artificial Intelligence Strategy 2021-2025	Presidency of the Republic of Turkey's Digital Transformation Office (CBDDO), and Ministry of Industry and Technology (MoIT)	Makes auditing datasets a cornerstone of its approach to "demographic, cultural, social diversity and inclusiveness as well as international human rights law, standards, and principles" throughout the lifecycle of AI systems. Recommends auditing, alongside impact analysis, risk assessments, and other forms of traceability as means to ensure privacy, responsibility, and accountability in AI system use. States that "[t]echnical, procedural and educational tools," "audit guides," and "application-based technical audits" will be developed. This extends to the use of AI in public institutions, which will be the object of AI audits [125, p. 59, 73].
Uruguay	Artificial Intelligence Strategy for the Digital Government	Digital Government Agency (Agesic)	Recommends "audits by impartial parties" for AI systems and related decision-making algorithms used in public administration [126, p.7].

Source: OECD's live repository of AI strategies and policies and IPIE analysis based on literature review [127].

ACKNOWLEDGMENTS

Contributors

Drafting Authors: 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), Sebastián Valenzuela (IPIE Chief Science Officer, Chile), Philip Howard (IPIE President and CEO, Canada/United Kingdom). Independent methodology reviews: Danaë Metaxa (USA) and Jack Bandy (USA/Australia). We are happy to acknowledge feedback from the Digital Democracies Institute and IPIE affiliates Luis Adrián Castro-Quiroa (Mexico), Luiz Valério de Paula Trindade (Brazil), Gabor Erdelyi (Hungary), Olivia Erdelyi (Hungary), Joan Hassan (Nigeria), Naoise McNally (Ireland), Carlos R. S. Milani (Brazil), Pradeep Nair (India), Fredrick Ogenga (Kenya), Zizi Papacharissi (Greece/USA). Copyediting: Beverley Sykes and Romilly Golding. We gratefully acknowledge support from the IPIE Secretariat: Egerton Neto, Donna Seymour, and Alex Young.

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, Simons Foundation, Swiss Re 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. Additionally, the Andrew W. Mellon Foundation also provided funding support to this project.

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 precis of the state of knowledge 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, 2024. *Global Approaches to Auditing Artificial Intelligence: A Literature Review*. SR2024.1. Zurich, Switzerland: IPIE.

DOI

https://IPIE.info/10.61452/BWYM7397

Copyright Information



This work is licensed under an Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

ABOUT THE IPIE

The International Panel on the Information Environment (IPIE) is an independent and global science organization committed to providing the most actionable scientific knowledge about threats to the world's information environment. Based in Switzerland, the mission of the IPIE is to provide policymakers, industry, and civil society with independent scientific assessments on the global information environment by organizing, evaluating, and elevating research, with the broad aim of improving the global information environment. Hundreds of researchers from around the world 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

9 783039 830008 >

