

The emergence of artificial intelligence ethics auditing

Daniel S Schiff^{1,†} , Stephanie Kelley^{2,†}
and Javier Camacho Ibáñez^{3,†}

Big Data & Society
October–December: 1–16
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517241299732
journals.sagepub.com/home/bds



Abstract

The emerging ecosystem of artificial intelligence (AI) ethics and governance auditing has grown rapidly in recent years in anticipation of impending regulatory efforts that encourage both internal and external auditing. Yet, there is limited understanding of this evolving landscape. We conduct an interview-based study of 34 individuals in the AI ethics auditing ecosystem across seven countries to examine the motivations, key auditing activities, and challenges associated with AI ethics auditing in the private sector. We find that AI ethics audits follow financial auditing stages, but tend to lack robust stakeholder involvement, measurement of success, and external reporting. Audits are hyper-focused on technically oriented AI ethics principles of bias, privacy, and explainability, to the exclusion of other principles and socio-technical approaches, reflecting a regulatory emphasis on technical risk management. Auditors face challenges, including competing demands across interdisciplinary functions, firm resource and staffing constraints, lack of technical and data infrastructure to enable auditing, and significant ambiguity in interpreting regulations and standards given limited (or absent) best practices and tractable regulatory guidance. Despite these roadblocks, AI ethics and governance auditors are playing a critical role in the early ecosystem: building auditing frameworks, interpreting regulations, curating practices, and sharing learnings with auditees, regulators, and other stakeholders.

Keywords

AI ethics auditing, algorithmic governance, semi-structured interviews, AI risk management, digital ethics, responsible AI

The emerging AI ethics audit ecosystem

Artificial intelligence (AI) is increasingly being used to aid decision-making in organizations. According to a prominent definition by the Organisation for Economic Co-operation and Development (OECD, 2024) adapted by the European Union (EU) in the recently enacted AI Act, an AI system refers to

a machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

Yet, the benefits of the technology are often accompanied by ethical issues, such as algorithmic discrimination (Buolamwini and Gebru, 2018), privacy breaches (Lacznia and Murphy, 2006), unclear human responsibility and accountability (Johnson, 2015; Martin, 2018), and a lack of transparency in the decision-making process (Ananny and

Crawford, 2016). These ethical issues are widely considered a management concern (Khalil, 1993). Consequently, firms and public sector organizations, specifically their managers, should bear ethical responsibility (Martin, 2018).

In 2016, as reports of unethical AI outcomes became more frequent, organizations developed AI principles—normative documents stating how AI systems should and should not be used by a firm's employees (Kelley, 2022) in hopes of articulating ethical priorities and minimizing

¹Department of Political Science, Purdue University, West Lafayette, IN, USA

²Sobey School of Business, Saint Mary's University, Halifax, Nova Scotia, Canada

³Departamento de Economía y Empresa, Universidad Europea Madrid, Madrid, Spain

[†]These authors contributed equally.

Corresponding author:

Daniel S Schiff, Department of Political Science, Purdue University, 110 N. University St, West Lafayette, IN 47907, USA.
Email: dschiff@purdue.edu



ethical risks related to AI (Schiff et al., 2021). However, AI principles are not enough to prevent unethical outcomes (Camacho and Villas, 2022; Munn, 2022; Rességuier and Rodrigues, 2020; van Maanen, 2022). The limits of ethics principles are well-known to business ethics experts (Kaptein, 1998; Metzger et al., 1993) who support the use of audits, among other tools, to drive more robust accountability and assurance.

More recently, in response to the explosion of generative AI models in the 2020s, prominent AI stakeholders have called for slowing or pausing the development of AI systems with human-competitive intelligence. These calls reiterate the need for improved risk management protocols that are rigorously audited by both internal and external experts. This support for AI ethics audits builds upon the mandated conformity audits of high-risk AI systems and voluntary codes of practice called for in the recently enacted EU AI Act. Researchers have likewise recommended AI ethics audits as a governance mechanism to operationalize AI ethics principles and prevent unethical AI system outcomes (Brown et al., 2021; Mökander and Floridi, 2022).

AI ethics audits, also referred to as “AI audits” or “algorithmic audits,” have been defined as

a process through which an [AI system] is evaluated [by] an AI auditor, [who] evaluates according to a specific set of criteria and provides findings and recommendations to the auditee, to the public, and/or to another actor, such as a regulatory agency, or as evidence in a legal proceeding. (Costanza-Chock et al., 2022: 1)

With limited regulation to date, evaluation criteria are often “ethics-based” and can include a firm’s AI principles, industry standards, best practices, guidelines, or other documents that specify criteria aligned with or over and above existing regulation (Mökander and Floridi, 2022). (While a variety of related terms are used loosely to refer to these audits, including governance audits, impact or conformity assessments, and assurance, we use the term “AI ethics audits” for simplicity.)

The development of AI ethics audits has been accompanied by an emerging AI ethics auditing ecosystem. It includes internal and external auditors (e.g., both start-ups and the Big Four accounting firms), auditing frameworks, risk and impact assessments, efforts by standards-setting organizations (e.g., IEEE, ISO, CEN-CENELEC), software as a service (SaaS) providers, and non-profits working on developing auditing criteria and certifications, as well as work on associated regulation. Some examples from this ecosystem include the ForHumanity Independent Audit of AI Systems,¹ the Government of Canada’s Algorithmic Impact Assessment,² IEEE CertifAIEDTM,³ and the International Association of Algorithmic Auditors,⁴ along with regulations such as the AI Act in the European

Union and New York City Local Law 144 on Automated Employment Decision Tools. These initiatives and other draft AI regulations commonly propose the use of AI ethics or governance audits, fueling the growth of the audit ecosystem.⁵

While support for AI ethics audits has been strong from academics and regulators, industry practitioners and other scholars have highlighted the lack of guidance provided. Numerous aspects regarding both the descriptive status and the normative ideal remain unclear, including an AI ethics audit’s scope, which activities constitute an AI ethics audit, the role of internal versus external auditors, what kinds of information sharing and reporting are required, the challenges of conducting audits, which stakeholders are engaged in the process, and how audits fit with existing AI ethics implementation initiatives and sectoral regulations (Costanza-Chock et al., 2022). Beyond the proposed role of AI ethics audits in AI system regulation and firm governance, the motivations for engaging in AI ethics audits are also unknown; for instance, whether audits are being implemented to reduce risk or for less altruistic reasons like ethics-washing (Bietti, 2020). Thus, despite the critical role of AI ethics auditing in AI governance, there remains very little empirical evidence on what constitutes an AI ethics audit, much less an effective one.

Developing an AI ethics auditing framework

Given these unknowns, we develop the following research questions in an effort to offer a descriptive framework and account of (early) AI ethics auditing: (1) What is the typical scope of an AI ethics audit?; (2) What drives auditees to request AI ethics auditing services?; (3) What processes, stakeholders, tools, frameworks, and deliverables are leveraged in an AI ethics audit?; (4) How is the success of an AI ethics audit measured?; and (5) What are the challenges associated with conducting an AI ethics audit?

To answer these questions, our study empirically evaluates the practices and perceptions of 34 AI ethics auditors largely employed in the private sector and engaged in private sector audits through the qualitative analysis of semi-structured interviews. The work builds on a few prior studies on AI ethics auditing by presenting interview results from a larger set of AI ethics auditors, both internal and external to companies, across several firms, industries, and seven countries. This allows us to provide a more comprehensive understanding of AI ethics auditing activities in practice, helping make prior framework-building efforts more complete.

In summary, we find that AI ethics audits follow the same process as financial audits, but tend to lack robust stakeholder involvement, measurement of success, and external reporting. AI ethics audits focus narrowly on technical AI ethics principles like bias, privacy, and explainability, reflecting regulatory emphasis on technical risk

management, unfortunately at the expense of less technical, but equally important principles. Challenges faced by the auditors conducting the work include interdisciplinary and cross-functional coordination, resource constraints, insufficient technical infrastructure, and ambiguous regulations. Despite these issues, AI ethics auditors play a crucial role in developing the AI ethics ecosystem by creating auditing frameworks, interpreting regulations, curating practices, and sharing insights with stakeholders.

After discussing the relevant literature and gaps in the next section, the paper presents the qualitative methodology used to conduct the research. This is followed by results: first, an account of the AI ethics auditing framework as it is beginning to manifest, which we find follows the process flow set forth by financial auditing: planning, performing, and reporting an audit. The discussion of second-order findings from the thematic analysis and the conclusions follow.

Situating AI ethics audits in the literature

Audits are a tool used to determine, through the examination of process or activity evidence, whether an organization is demonstrating the behaviors set out in some standard, which could include formal regulation, industry standards, or internal management metrics (Financial Reporting Council, 2020). They can act as a way for a firm to demonstrate adequate and meaningful accountability and should produce an economic or social benefit (Flint, 1988).

Financial audits

The most well-developed and practiced type of audit is the financial audit (Kaptein, 1998): the process of examining a company's financial records to ensure they are accurate and comply with laws and regulations (Porter et al., 2014). Financial audits are conducted by two groups: external and internal auditors. External auditing is an objective examination of an organization's financial statements by an individual or organization not employed by the organization being audited to determine whether the statements are free from material misstatement either due to fraud or errors (Porter et al., 2014). Internal auditing involves the same objective examination of financial documentation, but is run by the organization itself usually by the internal audit function, who also conduct an array of risk assessments, control assurance, and compliance work that maps to corporate governance (Gramling et al., 2004).

External and internal auditing follow a consistent framework that consists of planning, performing, and reporting (Porter et al., 2014). Planning involves understanding the environment being audited and conducting an initial risk assessment. Performing the audit follows, which includes scoping and evidence collection. Reporting closes the

audit loop and includes the provision of a report (usually to the Board of Directors and other stakeholders). The responsibilities of both external and internal auditing functions during these three stages are clearly defined in financial regulation: for example, the Generally Accepted Accounting Principles (Financial Reporting Council, 2020). In our analysis, we evaluate whether AI ethics audits follow a similar framework and discuss the implications for the process and prospects of AI ethics audits.

A key principle of both external and internal audits is that they are independent, meaning they operate separately from the department (or organization) being audited to ensure the provision of objective review. The framework for the Generally Accepted Auditing Standards outlines the fundamental principles of audits as follows: adequate technical training, adequate planning and staff supervision, sufficient understanding of the internal control environment to allow for competent evidence gathering, clear reporting versus standards, information disclosure, and independence (as noted above). Additional components outside these fundamental principles that have been found to support the quality of (external) audits are personal credibility, independence, openness of reporting, industry expertise, loyalty to minority shareholders, and professional skepticism (Warming-Rasmussen and Jensen, 1998). We build upon this well-developed classification of auditor types by leveraging the language of "internal" and "external" auditors when discussing AI ethics auditors within our study.

Business ethics audits

Beyond financial audits, business ethics audits are another common auditing activity that refers to "a systematic approach which makes a description, analysis, and evaluation of the relevant aspects of the ethics of a corporation" (Kaptein, 1998: 52). They have also been defined as "regular, complete and documented measurements of compliance with the company's published [ethics] policies and procedures," such as codes of ethics or codes of conduct (Rosthorn, 2000). Business ethics audits have several benefits for organizations. They can highlight weaknesses that could lead to unethical conduct (Madsen, 1990), determine whether employees take their ethical responsibilities seriously (Laczniak and Murphy, 1991), demonstrate that ethics programs are not just symbolic (Weaver et al., 1999), and help monitor unethical behavior (Maclean and Behman, 2010). Unlike financial audits, business ethics audits are often not released to the public (Kaptein, 1998), as they are designed to improve the ethical behavior of the organization's employees and not as a transparency tool for stakeholders; per these objectives, they are often conducted by internal auditors. Notably, however, the academic literature related to these audits struggles from a lack of cohesive theory (Kaptein, 1998). Our study presents a framework for an AI ethics audit, which we show is

related to business ethics audits in that AI ethics audits often focus on compliance with a code or framework of AI ethics (as opposed to just a code of ethics). We then discuss how the evolution of AI ethics regulation is leading to the development of AI ethics audits that more closely align with financial audits, as opposed to business ethics audits.

Impact and risk assessments

Risk and other assessments are closely related to, although distinct from, audits. Risk (also referred to as impact) assessments are “a type of fact-finding and evaluation that precedes or accompanies research, or the production of artefacts and systems, according to specified criteria” that are deemed relevant by stakeholders (Raab, 2020: 6–7). Risk assessments are used across a variety of industries and business units and include assessments related to environmental, occupational health and safety, financial, operational, reputational, quality, information technology, privacy, cybersecurity, and other sectors and functions. While the term ‘risk assessment’ (or ‘impact assessment’) is often used interchangeably with the term ‘audit’ (including by our interviewees), risk assessments are considered just one component of a financial audit, part of the planning stage (Porter et al., 2014). We take care to differentiate between AI ethics audits and early-stage AI ethics risk assessments in our content analysis and discussion of results.

AI ethics audits and assessments

While AI ethics audits (and assessments) are a relatively nascent development, much can be learned from leveraging existing auditing and assessment tools (Raji et al., 2020; Rismani et al., 2023). Following this perspective, much of the research to date has focused on reviewing existing auditing frameworks and recommending what aspects should be used to create successful AI ethics audits or assurance processes (see columns 1 and 2 in Figure 1). The literature has also discussed many potential challenges of AI ethics auditing (see column 3 in Figure 1); however, there has been very limited empirical work validating these studies in real-world settings (see column 4 in Figure 1 for a summary of methodologies in past studies).

Beyond conceptions of AI ethics auditing as a whole, scholars and policymakers have also proposed assessments for both specific and overarching AI ethics issues, with a particular focus on fairness and bias (Landers and Behrend 2022; Saleiro et al., 2019), human rights (ECNL and Data and Society, 2021), and well-being impact-focused assessments (Schiff et al., 2020). See Ayling and Chapman (2022) and Costanza-Chock et al. (2022) for a review of existing assessment tools.

Building on the insights from the literature review, we now outline the methodology used to investigate the motivations, key auditing activities, and challenges of AI ethics auditing.

Literature	Attributes of successful AI ethics audits	Challenges of AI ethics audits include	Methods
Raji et al. (2020)	follow a consistent auditing process	limited documentation due to agile development methods, system complexity, limited traceability of model output, lack of reliable standards, and ability to anticipate the larger societal impacts of AI systems	 review
Clavell et al. (2020)	collaborative (between auditors and AI developers)	limited access to the protected attributes (e.g., race, gender) necessary to conduct bias and discrimination risk assessments	 case study
Mökander & Floridi (2021)	continuously monitored, systems-based, question-based, aligned with organizational goals, focused on continuous improvement	lack of clarity on who audits whom, AI system access issues, harmonizing standards across a decentralized organization	 review
Mökander et al. (2021)	(in addition to attributes from Mökander & Floridi (2021)): traceable, accountable	see Mökander & Floridi (2021) challenges	 review
Mökander & Floridi (2022)	integrated into product development, integrated into existing governance structures, risk-based, based on a set of AI principles, measurable	harmonizing standards across a decentralized organization, determining scope, change management and communication, and designing audit that can handle external and internally developed AI systems	 case study
Costanza-Chock et al. (2022)	publicly disclosed, consider real-world harm, involve stakeholders, follow standards, be accredited	a lack of standards, limited regulation, auditee buy-in, lack of incident reporting by employees; challenges may differ between external and internal audits	 interview, survey

Figure 1. Conceptions of AI ethics audits in prior literature.

Methodology

We used a semi-structured interview approach, followed by a directed content analysis (Hsieh and Shannon, 2005) complemented by a secondary thematic analysis (Nowell et al., 2017) as our primary methodologies. Qualitative research methodologies have been used extensively to study auditing in practice (Hazgui and Brivot, 2020; Power and Gendron, 2015) and have been recommended as a critical methodology for emergent business ethics research (Reinecke et al., 2016). The methodology is a strong fit given our interest in the newly developing practice of AI ethics auditing in light of significant ambiguity, complexity, and socio-technical aspects that beg deeper elaboration. Below, we present details of our study recruitment process and sample before describing the qualitative analysis approach.

Recruitment and sample summary of AI ethics auditors

We identified the relevant population for this study as ‘AI ethics auditors’ defined as individuals who play a role in coordinating, facilitating, or conducting reviews of an organization’s AI governance or ethics approach. The research team conducted an initial search based on a purposive, selective sampling strategy via LinkedIn, organization websites, and interviewers’ personal networks to identify an initial set of 50 AI ethics auditors. These individuals were invited by the researchers to participate in April and May of 2022, resulting in 18 first-round interviews. In June, July, and August of 2022, the researchers invited 50 additional AI ethics auditors to participate, resulting in an additional 16 participants. Overall, a total of 34 of 100 individuals contacted were interviewed via 26 individual and three group interviews across 23 unique organizations from seven countries.⁶

Notably, our respondents all come from the private sector, although some have academic or civil society affiliations, as well. Furthermore, while a few individuals engaged in audits of government AI systems, the vast majority of the work pertains to the private sector context. While this was broadly reflective of our understanding of the core ecosystem, it nevertheless has implications for the sample, tenor of findings, and generalizability of results to public or non-governmental sector AI ethics auditing, such as by media and civil society organizations that may function as critical ‘watch dogs.’ Additional research is needed to contrast auditing across these contexts.

As per our inclusion criteria, we incorporated both individuals inside of an organization responsible for overseeing AI ethics audits internally (internal or first-party auditors) and individuals outside of a given organization (external, primarily second-party auditors). The individuals worked in both established companies and relatively new start-ups focused on AI ethics auditing and governance. Internal auditors were employed in financial services, telecommunications, retail,

and insurance, while external auditors were employed at consulting firms, software-focused start-ups, and law firms. Interviewees held a wide variety of functional roles across industry sectors, with titles, such as President or CEO, Product Manager, Director of Analytics, Researcher, and Chief Compliance Officer, among others. We deliberately sought diversity in terms of firm size, age, location, sector, as well as interviewee functional role and gender.

Table 1 provides basic descriptive information about the sample presented primarily in terms of participating organization characteristics. We also report participant gender.⁷

Semi-structured interview approach

We carried out individual and group semi-structured interviews during the months of April through August 2022 based on a semi-structured interview protocol iteratively developed by the research team and pilot-tested. The study was approved by the ethics boards at each author’s institution at the time of the research (approval numbers: H22108, GBUS-744-22, 2022/28). All interviewees provided appropriate region-specific consent and did not receive compensation.

The interview instrument was developed in anticipation of a directed content analysis approach (Hsieh and Shannon, 2005) using existing literature on AI ethics audits and non-AI audits and assessments (Carcello et al., 1992; Stoel et al., 2012), as discussed in the literature review. The interview script captured the following domains: (1) organization auditing scope and planning process; (2) motivations to engage in auditing; (3) auditing activities and outcomes; (4) engaged stakeholders; (5) alignment with and attitudes toward ethical concepts; (6) organizational and technical challenges; and (7) recommendations. Part A in the Online Supplement provides the full semi-structured interview protocol.

Interviews were conducted in English or Spanish and took approximately 45 min. Interviews were recorded, transcribed using the natural language processing tool Otter.ai, translated into English by the research team when appropriate, and then reviewed for accuracy. Due to the nature of the content, which could be sensitive to organizations or their clients, we allowed all participants to provide redactions or suggest other changes.

Coding and analysis

Following the interview process, the research team first followed a directed content analysis method, which included the development of codes both before and during the initial analysis. Codes were derived from the research questions, interview protocol, and related literature, and an a priori codebook was created. Additional codes and associated definitions were added, and others were refined across several iterations of early data analysis.

Table 1. Summary of participating organizations ($n = 23$) and interviewees ($n = 34$).

Variable	Count (%)	Variable	Count (%)
<i>Primary work location</i>		<i>Company sector</i>	
United States	9 (39.1)	AI ethics consultancy	7 (30.5)
Spain	6 (26.0)	Software	5 (21.7)
Canada	3 (13.0)	Legal	3 (13.0)
United Kingdom	2 (8.7)	Consulting	3 (13.0)
Australia	1 (4.4)	Financial services	2 (8.6)
India	1 (4.4)	Telecommunications	1 (4.4)
Singapore	1 (4.4)	Retail	1 (4.4)
		Insurance	1 (4.4)
<i>Internal or external auditor</i>		<i>Company age</i>	
External auditor	18 (78.3)	Start-up	13 (56.5)
Internal auditing role	5 (21.7)	Established firm	10 (43.5)
<i>Company scope</i>		<i>Interviewee gender ($n = 34$)</i>	
Governance audits	10 (43.5)	Man	22 (64.7)
Algorithmic audits	5 (21.7)	Woman	12 (35.3)
Software-as-a-service	4 (17.4)	Non-binary	0 (0.0)
Governance and algorithmic audits	4 (17.4)		

The researchers conducted a coding check by separately reviewing two individual interviews each using the preliminary codebook. After several rounds of iteration, the team engaged in a synchronous coding exercise to test the codebook. Finally, the two researchers who conducted the full coding of the 34 interviews engaged in a round of asynchronous coding before agreeing on the final codebook consisting of 96 codes. The combination of peer debriefing, synchronous, and asynchronous coding among the researchers helped to establish reliability (Hsieh and Shannon, 2005). The two coders divided the interviews randomly and coded them separately using the codebook⁸ and qualitative coding software ATLAS.ti (22).

Following the directed content analysis, the coders engaged in a round of thematic analysis primarily at the level of code domains as a means of identifying and reporting higher-level themes found in the interview data. Following the thematic analysis methodology set forth by Nowell et al. (2017), we triangulated prominent themes, diagrammed them to make connections, sometimes engaging in a process of theoretical redescription, all while keeping detailed notes about their development. We discussed the themes among the researchers and revised them until we reached a consensus on the prominent themes that aligned with our research questions. We first report the findings of the directed content analysis, followed by the second-order thematic results in the key themes discussion section.

A framework for AI ethics auditing: planning, performing, and reporting

The findings that constitute the identified framework for AI ethics auditing are presented in line with the financial auditing approach of planning, performing, and reporting, which

we find effectively describes the process of AI ethics audits. We begin with a discussion of what drives a company to engage in AI ethics auditing, followed by key findings across planning, performing, and reporting, and a discussion of the challenges faced by AI ethics auditors.

Before the audit: regulation and reputation as drivers of engaging in an AI ethics audit

Regulatory requirements and reputational risk were the most common drivers for the adoption of AI ethics audits. The majority⁹ of interviewees discussed regulatory motives, followed by a plurality noting reputational ones. For instance, interviewees noted that adoption of AI ethics auditing will be: “...*mainly driven by regulation, because in the realities of today, companies have multiple priorities to abide by, and there are only so many priorities they can pursue*” (External, UK, Woman).

Another interviewee agreed that “*there’s probably no greater motivator than realizing regulatory oversight is not just on the horizon*” and emphasized that the European Union (EU) AI Act “*will be a massive driver*” (External, Canada, Man). The EU AI Act was by far the most common regulatory development mentioned, lending further credence to the likelihood that this regulation will facilitate an international harmonization of regulations (i.e., the Brussels Effect) (Bradford, 2020; Siegmann and Anderljung, 2022) as auditors and companies shape much of their thinking around this legislation.

Participants also mentioned a range of other existing or proposed regulations, standards, and ethical frameworks that play a role in their thinking and auditing activities. For example: at the international level, the European Commission’s Guidelines for Trustworthy AI; at the national

level, the United Kingdom's (UK) Algorithmic Transparency Standard and US National Institute of Standards and Technology (NIST) AI Risk Management Framework; at the subnational level, New York Local Law 114 on automated employment decisions; and at the sector-specific level, the US' SR117 on model risk management.¹⁰ AI regulation as an extension of existing data regulation—as proposed by Mökander et al. (2021)—was rarely discussed.

Even concerning regulatory motives, interviewees described variation in how seriously these drivers are taken. For instance, one interviewee noted that companies “*will not feel the heat for several years*” given that “*GDPR was passed seven years ago...and still, there's very little effort in that respect*” (External, USA, Woman). Another auditor noted that companies could take a “*reactive*” or “*proactive*” approach (External, USA, Woman), such that even the consensus view that responsiveness to emerging regulation was important was itself subject to significant flexibility. Nevertheless, a common view was that “*the most pressing and most recent kind of motivation...is now just compliance with upcoming regulations*” (External, USA, Man).

Reputational motives were the second most commonly mentioned¹¹ and often, but not always, associated with a reactive style of engagement, which was sometimes described as instrumental or minimalistic: “*The starting point which we've had is significant public backlash or outrage in response to some risk that was uncovered, or something that causes significant pressure from the outside onto your company*” (External, USA, Man). However, we found that even these relatively instrumental drivers were part of a web of interconnected and evolving motivations, including prosocial goals alongside economic rationalizations. Interviewees described how reputational motives, for instance, were associated with an emphasis on customer trust, employee trust, a desire for ethical culture surpassing regulatory requirements, and even recognition that proper AI ethics auditing is just essential for AI performance:

I think that in all the cases we have audited, two elements have come together: conviction, because they were convinced that artificial intelligence ... has to be ethical ... And then they also saw a reputational issue, a branding issue. (External, Spain, Man)

[Companies] don't want to get in trouble with either the media or the regulators, there's that kind of fear. But I would say that's not a sustainable thing. Even those who start off like that, what they quickly realize is, we don't want to do this only for reputational reasons, or regulatory reasons. Frankly, if our models are not good enough, we don't want to use them, because they're just not good. (External, Singapore, Man)

In contrast, there were fewer consensus about growing public concern or awareness by investors and corporate boards playing a major role.¹² Yet, in line with the importance of leadership buy-in as noted by Kelley (2022), participants also pointed to individuals as important drivers, typically CEOs and other organizational leaders who themselves might be influenced by multiple motivations.

[...] it's driven by certain personalities, as well. So, you have one or two senior leaders who believe in the importance of this. (Internal, Canada, Man)

Unless you change the mindset at the top, you're not going to stick. (External, UK, Man)

Planning an AI ethics audit: open-ended scope determination and limited stakeholder engagement

Most of the interviewees' organizations (18 out of 23) were external to the organization they were auditing and had previous experience in consulting, auditing, data, or quality management. The other participants were internal, employed by an organization, and responsible for conducting audits within that firm. External auditors offered auditing services to both large companies and start-ups in several sectors, with a focus on highly regulated domains: financial services, insurance, technology, healthcare, hospitality, and telecommunications.

Participants noted two distinct approaches to auditing: governance and algorithmic audits. Auditors (internal or external) who employed a governance approach typically focused on a larger set of AI systems, their development processes, and associated organization-level structures. Alternatively, some auditors took an algorithmic approach, where the audit was generally centered on the data, performance, and outcomes of one or more AI systems or algorithms, but not the surrounding processes, although some governance auditors incorporated algorithmic audits as well (see Table 1). Lastly, the software-as-a-service (SaaS) providers were all external to the audited organization and focused on providing technical tools and the associated servicing of those tools (usually via a subscription-based fee) to support AI ethics principle assessments (typically bias, privacy, or explainability assessments).

The definition of the scope of the audit worked out between auditors and their clients was, therefore, critically important according to participants because it had substantial impacts on the types of auditors, associated activities, and possible gaps. Other factors involved during the scoping stage included the planned duration of the audit (often open-ended) and the determination of stakeholders to be engaged (often focused narrowly on certain key stakeholders):

It is so important to first define the purpose of the audit, the objective I am pursuing, and the scope I want to cover. The levels of audit I want to get to, whether it is a technical model, statistical models or whatever, down to the data level. (External, Spain, Man)

The organizations and teams involved in the audits were often multidisciplinary, with experts in data science, ethics, data protection and privacy, compliance, and legal aspects, etc., involved depending on the agreed-upon scope. Each engagement typically had a lead individual managing the project who called in colleagues as necessary. The average duration for an audit, according to the interviewees, varied from a few weeks to a year or more depending on factors, such as client availability, availability of necessary data and evidence, or requirements for conducting additional fieldwork. Overall, the scope in terms of duration, activities, and deliverables was often highly contextual and negotiated in an ongoing fashion, with external auditors hoping for continued engagements as part of their joint auditing and consulting activities.

We also asked auditors which stakeholders they considered important or engaged with as part of the auditing process. Figure 2 depicts the frequencies of the most commonly mentioned stakeholders. Auditors were substantially more likely to engage with technical individuals and teams, such as technology and data strategy leads, data scientists, and ML engineers. In certain sectors, such as finance, individuals in risk management, compliance, and legal settings were among the key stakeholders while involvement with laterally relevant functions like data protection offices varied. Executives and other business and product leads were also common stakeholders, including system and product design and development teams.

One associated complication was that auditors were solicited by, and interacted with, different departments with potentially competing interests, such as data scientists, executives, or risk professionals. Navigating between

these different teams and finding a common language was a key challenge:

You definitely could see there is a friction between different teams. Data science wants to be technical, they don't necessarily want to hear opinions of other teams. [...] For me, the success in that project was to actually [...] forcing them to have that dialogue with each other [...] to translate their technical thoughts into a language that the compliance team or the management team would be able to understand. (External, USA, Woman)

However, the results also revealed that stakeholders, such as the general public and vulnerable groups (as well as shareholders), are far much less likely to be engaged compared to core technical and risk professionals. This contrasts with the near-ubiquitous advice in scholarship and industry to support diverse and public engagement (Buhmann and Fieseler, 2023) and reveals significant gaps in auditing practice.

Performing an AI ethics audit: emphasis on risk identification and model validation

Regarding the core activities that constitute most audits, two activities stood out beyond the rest: risk management and model validation. The risk management approach heavily focused on risk identification with less robust efforts on assessment (i.e., measurement) much less mitigation, and employed various tools ranging from scorecards to questionnaires to identify operational, ethical, legal, and reputational risks. Overall, the vast majority of audits followed a risk-based approach despite the concern raised by Raji et al. (2020) that this approach may not be well-suited to anticipate the larger societal impacts of AI systems, involving complex human–AI interactions and feedback effects. Participants also referenced the need to document “*evidence for identification, prioritization, and maybe also justification for why [a] certain risk was identified or evaluated as low or high*” (External, USA, Man).

Before launching the product to the market, they need to go through an assessment process ... For instance, we have 14 ... very specific questions related to each of the five principles that we define. (Internal, Spain, Man)

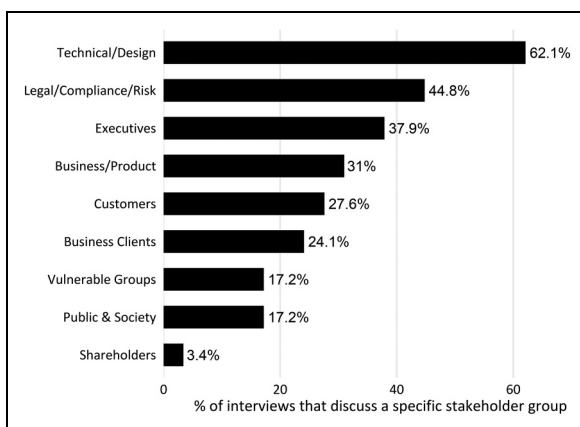


Figure 2. Stakeholders discussed by AI ethics auditors.

With respect to the other most common activity, model building and validation, this could involve auditing many aspects, from evaluating the design procedures and data collection approach to testing and validation of current models to assess bias, explainability, data drift, etc. Disparate impact analysis and other forms of algorithmic fairness testing were the most common. Several auditors were also tasked with building their own models to improve model

performance and/or bias metrics, while the decision about which thresholds or models were ultimately used given possible trade-offs between performance and fairness was left up to the parties who requested the audit.

There are three key findings related to this core activity: first, whether robust model validation occurred was dependent on whether the model and data were shareable or shared, often contingent on adequate governance and infrastructure; second, model validation was at times integrated into a broader analysis of the socio-technical system and surrounding development and governance processes, but more often limited to the assessment of model outputs; and third and relatedly, the scope of the validation sought by clients was often shaped by regulation, reinforcing the idea that regulation is not just a key driver of adopting an audit, but also for determining the objectives and activities of the audit.

So, we would examine the entire lifecycle of the model, from the design stage: did the developer choose the right approach of all the possible approaches? ... Did they select the right data? ... Are there data errors, which are going to produce modeling errors downstream? Was the model tested properly? Do we have confidence that it actually works as intended? (Internal, Canada, Man)

Participants also noted that organizations using AI are increasingly aiming to advance their ability to monitor and track models after they have been deployed to evaluate if they are working as expected over time. As such, continuous post-deployment monitoring and associated controls were mentioned as essential tasks related to model validation. The relevant tasks thus not only went beyond initial assessments, but also involved establishing supervision mechanisms, such as dashboards and other visualizations. These allowed for manual review or automatic flagging related to key performance metrics, demographic biases, and model or data drift based on identified thresholds and associated deviations. However, identifying the appropriate controls, thresholds, and responses was non-obvious. Making these decisions was rendered even more difficult by the rapid evolution and uncertainty of regulation, as well as the wide variety of model techniques and use cases embedded in their respective social contexts.

Several of the systems that we have audited have stopped being used, so the system just doesn't exist anymore, so we can't go back and audit them again. But in some cases, when we have been able to go back, what we are finding is that this space is evolving so quickly, that we can't really just use the metrics and the understandings that we developed the year before. (External, USA, Woman)

Reporting an AI ethics audit: less clarity on measures of success and limited external reporting

Given that a major need for AI ethics auditing as a new field is evaluating what constitutes a quality audit, we asked interviewees how—or indeed whether—they measured success. Interviewee answers related to two main areas: quantitative indicators and perceptions of organizational impact and change. Regarding indicators, many auditors discussed improvements to certain OKRs or KPIs, especially reducing disparate impact and improving model accuracy, as well as occasionally profit indicators or related metrics (depending on sector, these included conversion rate, retention rate, time-to-market, revenue, etc.).

It's a really good feeling to be able to go back to a client and say, hey, we dropped these 20 features from the model, tweaked these hyper parameters, and now you've got a model that has far less disparate impact. (External, USA, Man)

Yet, many auditors noted they did not have specific metrics of success, and some were intrigued by the question as they had not thought it through in an explicit sense. In discussion, however, they identified that completing an audit report, fulfilling the initial scoped deliverables of the audit, improving organizational awareness, and especially witnessing improvements in organizational capacity and governance based on auditor recommendations were some of the most meaningful indications of impact.

What I can say is that the success starts when the recommendations are being implemented and applied, because in the end, what we really want is not to deliver a report if it is not useful for the client or if it is not going to change anything. I think success is about changing things in a positive direction. (External, Spain, Woman)

Furthermore, in almost every audit, auditors produced a technical report. External auditors, especially from consulting-focused organizations, typically created their own templates related to a variety of deliverables such as bespoke model validation reports and governance recommendation reports targeted at data scientists (e.g., integrated with Jupyter Notebooks) or business leaders. The final reports were almost entirely oriented toward internal audiences. Auditors did note a few examples where they suspected clients might elaborate or adjust a report to meet certain regulatory requirements, and some auditors designed their templates in anticipation of this. However, the practice of external reporting was mostly hypothetical and not in the control of auditors, suggesting many other reports currently function more as consulting artifacts and are not currently used to satisfy regulatory or public transparency goals. As one external auditor proposed:

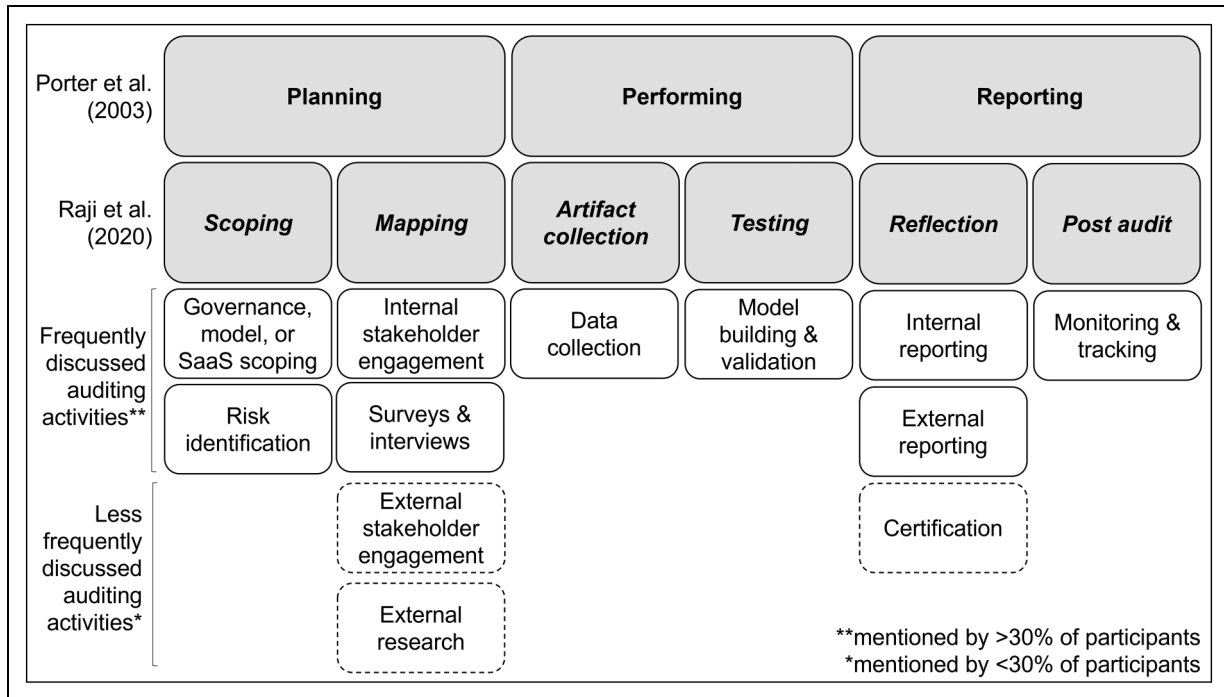


Figure 3. Activities discussed in AI ethics auditing aligned with existing auditing frameworks.

When you're reporting, you could potentially be providing a report just for internal use only, and then be editing it to create something that's much more public-facing like, on your website for transparency, for your customers, or to provide a nice to have or sweetener for regulators. (External, USA, Woman)

In summary, our qualitative findings evaluating the auditing process are largely commensurate with the high-level internal audit framework proposed by Raji et al. (2020) of scoping, mapping, artifact collection, testing, reflection, and a post-audit stage (see Figure 3). Our findings indicate that this framework is highly representative of how the AI ethics audit process is developing in practice for both internal (first-party) and external (second-party) auditors. Furthermore, this process aligns with the financial auditing framework of planning (scoping and mapping), performing (artifact collection and testing), and reporting (reflection and post-audit stage) (Porter et al., 2014: 8; Raji et al., 2020), an association we visually map in Figure 3 with individual activities listed as part of an updated account and framework of AI ethics auditing.

Yet, we do observe great variability in what is involved in each of these high-level steps and whether they are carried out robustly. Regarding variability, for instance, Raji et al. (2020) suggested specific artifacts that should be created by auditors at a given stage, such as a social impact assessment and a use case ethics review during the scoping stage. Our findings suggest that auditors are often not generating consistent artifacts from their audits, but

rather adapt their activities in each of these stages depending on the AI system being reviewed, the organizational engagement, and the objectives, even when auditors begin with an attempt at driving a standardized framework. Perhaps, more importantly, we witness less evidence of robust auditing associated with the reporting stage. For instance, while there are some advances in post-deployment monitoring, we see very little evidence of external reporting despite discussion of its importance. Further, the plurality of audits is oriented at evaluating a limited group of technical topics, representing a fairly narrow subset of the diverse socio-technical dimensions scholars associate with the AI lifecycle.

Across the audit: challenges identified by auditors

The most common challenges¹³ discussed surrounded uncertainty and ambiguity resulting from preliminary and piecemeal regulation and an associated lack of effective and vetted best practices. Other challenges included organizational complexity, interdisciplinary and cross-functional coordination, limited data availability and quality, a lack of baseline data and AI infrastructure, and an underdeveloped capacity of clients to even engage with auditors.

Foremost, auditors noted that the regulatory ecosystem is immature, and that they cannot readily answer questions about how to interpret regulations, although they are asked to do so by clients or effectively required to as part of their efforts.

I think it's all very immature, it's all very new, nobody knows where this is gonna go in the future. I think that we're all waiting for regulation. (External, USA, Man)

Another challenge is lack of regulatory clarity. It's still unclear, many of these questions are still unsettled by regulators. Makes it very difficult, because somebody asks you, in your expert opinion, what should we do? And your answer is, basically, I don't know, regulators don't know. (External, USA, Man)

The lack of mature regulation also meant that some clients may not have felt the need to resource their AI ethics and governance work in a way that even *enables* high-quality engagement with auditors. As one auditor noted,

People are talking about AI ethics, but like, actual budgets allocated for AI ethics? Why is that not happening? The reason why that doesn't happen is there's no well-defined regulation. (External, USA, Man).

Along these lines, many auditors noted that there simply are not clear standardized tests or metrics regarding how to assess even common issues like algorithmic bias, which creates uncertainty, even when auditors' work requires making associated decisions. Similar concerns around a lack of viable standardization have been voiced previously by Mökander and Floridi (2021, 2022) and the practitioners surveyed by OpenLoop (2022). Auditors further worried that even commonly utilized practices may be insufficiently robust, leading to neglect of important social and ethical issues. This was especially true when auditing strategies were limited to technical or 'measurable' approaches such as algorithmic fairness auditing centered on statistical tests.

When I first got into the field, I thought, okay, here's what we're gonna do. We're going to define what we mean by fairness. And then we're gonna go out and measure it. And then we'll be able to say, this one's fair. That one's not fair. After I did some reading, I abandoned this idea. (Internal, Canada, Man)

I'm a little bit wary that this sort of sensationalizing and promotion of these technologies as if they can solve the governance problem is a little bit negligent or reckless. I think that it gives this false illusion that we can engineer our way out of these problems. And that's just not the case. (External, USA, Woman)

As suggested by Mökander and Floridi (2021), simply getting sufficient access to the AI systems is itself a significant auditing challenge. Our findings expand on this, as auditors indicated that many companies lack robust data and model governance and documentation, meaning it is

difficult to find where data exist, how they were collected, what data were used by different models, and so on. This spells problems for identifying the appropriateness of data and models, understanding limitations and biases, and having access to basic demographic data to engage in techniques like fairness testing. The same is true for general processes around model design and evaluation, as auditors described, reflecting a lack of basic model inventories, provenance for modeling decisions, or data traceability.

In terms of challenges, I think the issue of data, it's one of the most important ones. Even sometimes this data is not well structured or is not only a problem of access and property rights that also, but it's a matter of even how code is written or how the data sets are built. (External, Spain, Man)

It also exemplifies how weak the legs of this whole technical ecosystem around us are. Because no one can locate the data or the people that came up with trade-offs. It's just really hard for them to understand how their own systems are working, which is terrible, scary. (External, USA, Woman)

These governance gaps meant the auditors spent much of their time trying to encourage clients to build basic data and model documentation and governance infrastructure. Furthermore and critically, we found that auditors may not only lack access to AI systems and data, but also to appropriate individuals and sufficient information to answer their questions in a timely fashion (or at all). The associated set of challenges included navigating organizational complexity, identifying the appropriate point people—a concern also raised by Mökander and Floridi (2021)—and keeping track of dynamic structures and processes. As auditors observed:

A practical limitation is getting to all the information in an organization. Some organizations are massive, and they are not well documented and the processes internal to the organization are not like you can just read a book on it. [...] If we can't embed ourselves in there 100%, then it's going to be hard for us to really come up with the right solution. (External, USA, Man)

You can't just look at a product in isolation, without looking at how it works in the wider organization, about what kinds of employees they have, their capacity, their competence, the resources and budget behind them. Because it's all well and good having someone designated to in your data science department to check for algorithmic bias, but if they've got no budget to go and do that, or if they've got no tools to assist them, how can they realistically go ahead and do that? (External, USA, Woman)

Coordinating across multiple different teams was another common difficulty as individuals with diverse functional roles at times had competing perspectives and priorities. Interviewees noted that some employees appeared to hold objectives that differed from those of other colleagues even within their own companies, leading to lack of coordination, communication, and even resistance, as noted below.

That's part of the challenge that a lot of these organizations have, three teams working in silos on these issues, but none ... are speaking to one another. (External, USA, Woman)

The first one is a coordination challenge, because it is a conversation between multiple teams. Usually, this takes a very long time to get everyone [to] buy in and [be] trained and so on. So first off, you don't get anywhere without that. (External, Canada, Man)

In combination, this means that auditors must work across social, technical, and governance functions, but while often lacking effective access, buy-in, and resources across these functions to do so. Moreover, without standardized understanding of expectations or processes, either from regulation or from the companies themselves (who often do not know what they want or need), auditors are effectively asked to address a challenge that can only be solved by resolving broader organizational complexities and providing regulatory certainty. As auditors summarized:

I think a lot of people's view is that no organization is ready to put their AI algorithm and autonomous systems through an audit. (External, UK, Man)

Our clients are nowhere near ready to perform audits. Because in order to perform an audit, you need to have governance of AI. (External, UK, Woman)

Discussion: key themes in the emerging AI ethics audit ecosystem

Following our first-level findings, we also conducted a cross-domain analysis using underlying codes and compared results to the existing literature on AI ethics auditing. We identified several higher-level themes that may be especially important in understanding the status of the AI ethics auditing ecosystem today.

Ambiguity, but progress in the development of the AI ethics audit ecosystem

Perhaps foremost, our findings highlight how AI ethics auditing is a complex and challenging endeavor on several levels. Auditors are keenly aware of the ambiguity between auditing and consulting activities, and there is

some concern from participants about professional independence in the AI space as many explicitly noted that they may not be engaged in "true" auditing (Costanza-Chock et al., 2022). This ambiguity in the early development of AI ethics auditing is exacerbated by factors like emerging and piecemeal regulations, lack of harmonization around standards and best practices, an absence of measures for determining AI ethics auditing quality, and the associated lack of accredited auditing procedures and certifications. Additionally, as early-stage AI ethics auditing organizations work to build out their practice, they need to resource their organizations and satisfy the variable preferences and needs of diverse companies: all of these tasks require the audited organizations to develop their data and AI infrastructure and organizational capacity to a sufficient level to engage in audits in the first place (Clavell et al., 2020).

Despite these challenges, we do observe that the AI ethics auditing ecosystem has progressed on some fronts. Compared to Costanza-Chock et al. (2022), who noted that effectively no standards are referenced, we find that auditors mentioned dozens of international and subnational regulations, emerging standards, and theoretical frameworks (see Part B, Online Supplement for examples). There are indications of convergence around the EU AI Act and the US NIST AI Risk Management Framework.

Furthermore, our findings indicate that AI ethics auditors have had success in a diverse array of tasks. This includes operationalizing ambiguous regulations, improving model fairness given certain criteria, and even promoting socio-technical frameworks that may be unfamiliar to technical teams. Auditors have helped spur organizational change in terms of guiding advances in data infrastructure and model inventories, promoting goals like interdisciplinarity and traceability, developing and sometimes evaluating progress on ethical principles, and more.

Finally, a key takeaway is that AI ethics auditing is evolving along lines most closely connected to financial auditing, although it also has connections to business ethics auditing, as well as novel features and challenges. These core relationships are helpful in suggesting directions for theoretical and practical development and in cautioning about potential pitfalls.

Gaps in the AI ethics audit ecosystem: measuring success, reporting, and stakeholder engagement

Indeed, in several key respects, AI ethics audits have not yet realized practices called for in literature and across public, private, and NGO sectors. Three areas for continued progress are measuring outcomes and establishing success, effective and public reporting, and broader stakeholder engagement. For example, Mökander and Floridi (2022) and Raji et al. (2020) noted the importance of measuring outcomes, reflection, and a post-audit stage. Yet, we

found that very few AI ethics auditors had well-formulated conceptions or specific qualitative or quantitative criteria for what ‘success’ meant in the context of their audits, bar a couple of organizations developing AI ethics certification schema. For some, achieving a certain statistical threshold, minimizing model bias, completing a final report, or observing cultural change in an organization is a marker of success. Indeed, a number of interviewees were surprised and intrigued by the interview question on “how they determine audit success” and expressed interest in thinking more carefully about this in the future.

Second, we identified limited forms of stakeholder engagement, deviating from repeated guidance urging broader efforts. Auditors are variably solicited by and engage with business leadership, individuals in law and risk, and data science and design teams. Indeed, we found that auditors spend most of their time talking directly to technical or governance teams often to gather information on data or AI systems to support algorithmic bias assessments. In contrast, in only a few cases did auditors proactively reach out to broader stakeholders (e.g., conducting qualitative interviews with members of the public). This suggests a need for progress in realizing diverse, public, and interdisciplinary stakeholder participation, a call that has been made across governments, companies, standards organizations, civil society, and academia and an active area of research (Buhmann and Fieseler 2023; Schiff 2024).

Overall, while key regulatory documents like the EU AI Act and NIST Risk Management Framework call for public transparency, accountability, and interdisciplinary subject matter expertise, these processes are currently secondary at best in AI ethics auditing practice, especially regarding stakeholders external to firms. Resource limitations, perceived regulatory pressures, lack of clear best practices for stakeholder engagement, concerns about reputational risk, and even trade secrecy concerns could be limitations. Auditors and auditees may need to devote more attention to developing these capabilities and addressing the associated challenges if they are to satisfy this repeated demand for deep engagement. Yet, other forms of public engagement and accountability may still be realizable, such as through public access to transparent reporting, attention via civil society and media watchdogs, and political engagement.

How AI ethics auditors are shaping AI governance

While most external AI ethics auditing efforts do not (yet) meet the key criteria specified by Warming-Rasmussen and Jensen (1998) of independence, sufficient understanding of the internal control environment to enable competent evidence gathering, or openness of reporting, it is arguably premature to articulate this as a failure of the AI ethics auditing ecosystem. Instead, our results point to the

current situation as transitional and developmental. If the AI governance community ultimately fails to develop sufficient independent external auditing capabilities and relies largely on consulting efforts, this would indeed fail to meet standard criteria for successful auditing, as cautioned by Raji et al. (2020) and others. Yet, our interpretation is that AI ethics auditors who take on the risk in this early stage are playing a critical role in developing the broader ecosystem’s ability to transition to more robust auditing.

As noted previously, AI ethics auditors inherit numerous challenges, one being a patchwork of standards and forthcoming regulations. This means AI ethics auditors are often required to, explicitly or implicitly, determine which regulations and standards are relevant, make hard calls about disambiguating vague requirements into operationalizable frameworks, decide which technical or social tools are relevant, and co-develop processes with under-resourced or under-prepared organizations for implementing these decisions. Auditors thus must have broad familiarity with multi-layered regulations and interdisciplinary techniques and must engage with a variety of organizational stakeholders with potentially disparate perspectives and incentives. They must do so while also engaging in careful case-making and balancing acts related to company interests in regulatory, reputational, or financial goals, sometimes against ethical ones. As a result, we found that auditors often create their own frameworks, software packages, reporting templates, and other tools to operationalize AI ethics and governance, thereby playing a critical role as the interpreters and translators of the ecosystem.

Implications for practitioners and scholars

We offer a few practical implications based on the findings and explicit recommendations of interviewees. For organizations considering auditing, key goals include adequately resourcing AI governance efforts, building baseline technical and data infrastructure, identifying relevant point people and responsibilities, and coordinating contact with AI auditors along a streamlined process to maximize appropriate sharing of information and minimize internal dissensus. For auditors, strategies include considering governance-level audits for their greater robustness, tracking key emerging regulations for alignment, encouraging auditees to meet certain scope requirements for effective engagement, and working toward other goals like broader stakeholder engagement, external reporting, and more robust treatment of ethics. Both auditees and auditors can continue to play a key role in sharing best practices with actors across the ecosystem, including standard organizations, academics, and policymakers. Finally, policymakers were repeatedly recognized as key actors with the capacity to substantially shape this ecosystem. According to our interviewees, their efforts to develop consensus around sufficiently tractable and detailed recommendations and provide guidance that minimizes ambiguities are indispensable.

For scholars, in addition to the theoretical directions discussed above, future research could look to identify variations in AI ethics auditing as it evolves across different sectors, types of auditors and audits, countries with different cultural and policy environments, highly regulated vs. less-regulated industries, high-risk vs. lower-risk AI systems, and so on. Furthermore, as interviewees noted, there are currently relatively few efforts among AI ethics auditors to engage with members of the public, report externally, or even define auditing success during their work with clients or internally. As auditing standards and practices advance and become formalized through law and consensus, researchers will need to identify goals and challenges associated with numerous sub-components of AI ethics and governance auditing.

Conclusion

In this paper, we present an empirically grounded framework for AI ethics audits developed via a study of current AI ethics auditors, their perceptions, and their activities in practice. The study thus presents an alternative view to the primarily normative-based study of AI ethics audits and AI ethics in general and consists of the largest-scale interview study of AI ethics auditors to date. Drawing on English and Spanish language interviews of 34 auditors in seven countries, both internal and external to companies across sectors, we gather insights from individuals advancing AI ethics auditing and evaluate motivations, activities, and barriers to effective auditing practice. We present our results with respect to existing financial and business ethics audits, building upon the extensive practical frameworks developed there, and advancing prior framework-building efforts in the AI ethics and governance auditing space. We offer a descriptive account of the field and a tractable framework to enable future theoretical development.

Our research is subject to limitations common to qualitative research and studies of nascent and evolving fields. While our sample is relatively large and diverse compared to prior research, it still captures only a fraction of the broader AI ethics auditing community (only seven countries) and represents only a snapshot of a highly dynamic ecosystem. For example, our rationale and search approach led us to emphasize private sector organizations and auditors who overwhelmingly serve the private sector. Our study's results may not extrapolate as immediately to auditing in or by the public sector or civil society. Finally, given the variability of participants across multiple countries, we did not offer findings on possible regional or institutional differences within the scope of this paper nor comment on differences across individual auditor characteristics, such as gender; both research topics that should be explored in future work.

Overall, we find that the community focused on AI ethics auditing is rapidly evolving, experimenting, and

preparing for landmark changes in AI regulation, policy, and standards. We find that AI ethics auditors apply both governance and algorithmic auditing, serving clients with a complex and mixed set of motivations, with regulatory risk most salient, followed by reputational risk mitigation. While the auditing process often centers around processes like risk identification and overly centers technical model validation efforts surrounding bias or performance, the frameworks and strategies used are incredibly diverse and the chosen set of tools is unique to each auditor and auditee. We do observe though that a broader socio-technical and ethical approach is often valued by governance auditors, in part due to their expressed prosocial motivations.

Yet, auditors and the companies they serve face deep challenges related to interdisciplinary and cross-functional engagement, limited organizational capacity to engage in auditing, and a lack of technical and data infrastructure, as well as a broad absence of regulatory clarity and consensus on standards and best practices. In response to this ambiguity, auditors have served as ecosystem builders and translators, curating frameworks and building best practices, even as they socialize these ideas with clients, regulators, and other stakeholders. As recognized by auditors in this study, solving the challenges associated with AI ethics auditing will require a collective effort from auditors, companies, governments, academics, and beyond.

Acknowledgments

We are grateful to Jacqui Ayling for her support in research design and data collection. We also wish to thank all of our interviewees.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval and informed consent statements

The study was approved by the ethics boards at each author's institution (approval numbers: H22108, GBUS-744-22, 2022/2028). Participants provided verbal consent.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability

To promote confidentiality, interviewee identities and transcripts are not available. However, additional information on the research process is available via the Online Supplement or on request.

ORCID iDs

Daniel S. Schiff  <https://orcid.org/0000-0002-4376-7303>

Javier Camacho Ibáñez  <https://orcid.org/0000-0001-7565-5480>

Supplemental material

Supplemental material for this article is available online.

Notes

1. <https://forhumanity.center/independent-audit-of-ai-system>
2. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>
3. <https://engagestandards.ieee.org/ieeecertified.html>
4. <https://iaaa-algorithmicauditors.org>
5. For a more complete list of AI ethics ecosystem members, see: <https://www.eaidb.org>.
6. The final sample's regional composition in part reflects the home countries of the research team and the team's linguistic comfort (interviews were only conducted in English and Spanish). Team members initially reached out to organizations and individuals in their home countries to increase the diversity of the sample. However, the final sample also ultimately reflects our understanding of the prevalence of prominent AI auditing organizations and results from snowball sampling based on outreach to these organizations. For example, many major AI auditing organizations are headquartered in the US, Europe, and various high-income countries, as they tend to serve organizations in those regions, which also have more developed AI regulatory frameworks. Nevertheless, this can only provide partial insight into auditing ecosystems across a variety of national, regulatory, and institutional contexts.
7. While gender does not impact the results, the authors consider it important to include participants' gender in a traditionally non-diverse space (AI and STEM in general). Including gender in the participant quotes maintains awareness of the issue and promotes understanding of the implications of gender diversity and representation in STEM (e.g., in alignment with UN Sustainable Development Goal 5: Gender Equality).
8. The detailed codebook with the full list of the 96 individual codes and definitions is available in Part C of the Online Supplement.
9. For example, 104 out of 324 quotes related to motives mention regulatory ones, making this the most discussed driver.
10. A full list of regulations, standards, and ethical frameworks discussed is included in Part B of the Online Supplement.
11. Reflected in 69 of 324 quotes about motives.
12. The data collection was conducted before the prominent public release of LLMs like ChatGPT.
13. A total of 42 out of 161 quotes related to challenges mention regulatory limitations and uncertainty.

References

- Ananny M and Crawford K (2016) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20(3): 973–989.
- Ayling J and Chapman A (2022) Putting AI ethics to work: Are the tools fit for purpose? *AI Ethics* 2(3): 405–429.
- Bietti E (2020) From ethics washing to ethics bashing. In: Proceedings of ACM FAT* Conference (FAT* 2020), Barcelona, Spain. <https://doi.org/10.1145/3351095.3372860>
- Bradford A (2020) *The Brussels Effect: How the European Union Rules the World*. Oxford, England: Oxford University Press.
- Brown S, Davidovic J and Hasan A (2021) The algorithm audit: Scoring the algorithms that score us. *Big Data & Society* 8(10): –8.
- Buhmann A and Fieseler C (2023) Deep learning meets deep democracy: Deliberative governance and responsible innovation in artificial intelligence. *Business Ethics Quarterly* 33(1): 146–179.
- Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp.77–91.
- Camacho Ibáñez J and Villas Olmeda M (2022) Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI and Society* 37(4): 1663–1687.
- Carcello JV, Hermanson RH and McGrath NT (1992) Audit quality attributes: The perceptions of audit partners, preparers, and financial statement users. *Auditing: A Journal of Practice & Theory* 11(1): 1–15.
- Clavell GG, Zamorano MM, Castillo C, et al. Auditing algorithms: On lessons learned and the risks of data minimization. In: AAAI/ACM Conference on AI, Ethics, and Society (AIES'20), New York, NY, USA, 2020, February 7–8.
- Costanza-Chock S, Raji ID and Buolamwini J (2022, June 21–24) Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In: 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), Seoul, Republic of Korea.
- ECNL and Data & Society (2021) Recommendations for incorporating human rights into AI impact assessments. In: ECNL. The Hague, Netherlands: ECNL and Data & Society, pp.1–19. <https://ecnl.org/publications/recommendations-incorporating-human-rights-ai-impact-assessments>.
- Financial Reporting Council (2020) Auditors | Audit and Assurance | Standards and Guidance for Auditors. <https://www.frc.org.uk/auditors/auditassurance/standardsandguidance> (accessed 31 March 2023).
- Flint D (1988) *Philosophy and Principles of Auditing: An Introduction*. London, UK: Wiley.
- Gramling AA, Maletta MJ, Schneider A, et al. (2004) The role of the internal audit function in corporate governance: A synthesis of the extant internal auditing literature and directions for future research. *Journal of Accounting Literature* 23: 194.
- Hazgui M and Brivot M (2020) Debating ethics or risks? An exploratory study of audit partners' peer consultations about ethics. *Journal of Business Ethics* 175(4): 741–748.
- Hsieh H and Shannon SE (2005) Three approaches to qualitative content analysis. *Qualitative Health Research* 15(9): 1277–1288.
- Johnson DG (2015) Technology with no human responsibility? *Journal of Business Ethics* 127(4): 707–715.
- Kaptein M (1998) *Ethics Management: Auditing and Developing the Ethical Content of Organizations*. Dordrecht, Netherlands: Springer.

- Kelley S (2022) Employee perceptions of the effective adoption of AI principles. *Journal of Business Ethics* 178(4): 871–893.
- Khalil OEM (1993) Artificial decision-making and artificial ethics: A management concern. *Journal of Business Ethics* 12(4): 313–321.
- Laczniak GR and Murphy PE (1991) Fostering ethical marketing decisions. *Journal of Business Ethics* 10(4): 259–271.
- Laczniak GR and Murphy PE (2006) Marketing, consumers and technology. *Business Ethics Quarterly* 16(3): 313–321.
- Landers RN and Behrend TS (2022) Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist* 78(1): 36–49.
- MacLean TL and Behman M (2010) The dangers of decoupling: The relationship between compliance programs, legitimacy perceptions, and institutionalized misconduct. *Academy of Management Journal* 53(6): 1499–1520.
- Madsen P (1990) Managing ethics. *Executive Excellence* 7(12): 12.
- Martin K (2018) Ethical implications and accountability of algorithms. *Journal of Business Ethics* 160(4): 835–850.
- Metzger M, Dalton DR and Hill JW (1993) The organization of ethics and the ethics of organizations: The case for expanded organizational ethics audits. *Business Ethics Quarterly* 3(1): 27–43.
- Mökander J and Floridi L (2021) Ethics-based auditing to develop trustworthy AI. *Minds and Machines* 31(2): 323–327.
- Mökander J and Floridi L (2022) Operationalizing AI governance through ethics-based auditing: An industry case study. *AI and Ethics* 3(2): 451–468.
- Mökander J, Morley J, Taddeo M, et al. (2021) Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics* 27(44): 1–30.
- Munn L (2022) The uselessness of AI ethics. *AI and Ethics* 3(3): 869–877.
- Nowell LS, Norris JM, White DE, et al. (2017) Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods* 16(1): 1–13.
- OECD (2024) Explanatory memorandum on the updated OECD definition of an AI system. OECD Artificial Intelligence Papers, No. 8, OECD Publishing, Paris, <https://doi.org/10.1787/623da898-en>.
- OpenLoop (2022) Artificial Intelligence Act: A policy prototyping experiment operationalizing requirements Part 1. https://openloop.org/reports/2022/11/Artificial_Intelligence_Act_A_Policy_Prototyping_Experiment_Operationalizing_Reqs_Part1.pdf.
- Porter B, Simon J and Hatherly D (2014) *Principles of External Auditing* (4th ed.). Hoboken, NJ, USA: Wiley.
- Power MK and Gendron Y (2015) Qualitative research in auditing: A methodological roadmap. *Auditing: A Journal of Practice & Theory* 34(2): 147–165.
- Raab CD (2020) Information privacy, impact assessment, and the place of ethics. *Computer Law and Security Review* 37: 105404.
- Raji ID, Smart A, White R, et al. (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp.33–44. <https://doi.org/10.1145/3351095.3372873>.
- Reinecke J, Arnold D and Palazzo G (2016) Qualitative methods in business ethics, corporate responsibility, and sustainability research. *Business Ethics Quarterly* 26(4): 13–22.
- Rességuier A and Rodrigues R (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society* 7(2): –5.
- Rismani S, Shelby R, Smart A, et al. (2023) From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), pp.18.
- Rothorn J (2000) Business ethics auditing— More than a stakeholder's toy. *Journal of Business Ethics* 27(1/2): 9–19.
- Saleiro P, Kuester B, Hinkson L, et al. (2019) Aequitas: A bias and fairness audit toolkit. <http://arxiv.org/abs/1811.05577>.
- Schiff D, Ayesh A, Musikanski L, et al. (2020) IEEE 7010: A new standard for assessing the well-being implications of artificial intelligence. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp.2746–2753.
- Schiff D, Borenstein J, Laas K, et al. (2021) AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society* 2(1): 31–42.
- Schiff DS (2024) Framing contestation and public influence on policymakers: Evidence from US artificial intelligence policy discourse. *Policy and Society* 43(3): 255–288.
- Siegmann C and Anderljung M (2022) *The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market*. Oxford, UK: Centre for the Governance of AI. <https://www.governance.ai/research-paper/brussels-effect-ai>.
- Stoel D, Havelka D and Merhout J (2012) An analysis of attributes that impact information technology audit quality: A study of IT and financial audit practitioners. *International Journal of Accounting Information Systems* 13(1): 60–79.
- Van Maanen G (2022) AI ethics, ethics washing, and the need to politicize data ethics. *Digital Society* 1(9): 1–23.
- Warming-Rasmussen B and Jensen LH (1998) Quality dimensions in external audit services- an external user perspective. *European Accounting Review* 7(1): 65–82.
- Weaver GR, Treviño LK and Cochran PL (1999) Corporate ethics programs as control systems: Influences of executive commitment and environmental factors. *Academy of Management Journal* 42(1): 41–57.