Audit of AI Systems: Overview, Current Status, and Future Prospects

Andrew Grotto
Sagewood Global Strategies LLC
November 21, 2022

The AI audit ecosystem is immature, at best. Technical and other challenges are obstacles to generating audit results that relate to the harms people care about. Audit criteria for digital systems exist but how to apply them to AI systems is a work in progress. Without auditable criteria that map to settled expectations about AI system performance, an audit cannot purport to measure the system's performance against expectations. The training, accreditation, and professional ethics ecosystem for AI audit is in its infancy. Unrealistic expectations from policymakers and the public about the current potential for audit as a tool for reducing AI risk could result in misplaced confidence in audit results and undermine the credibility of audit generally.

Contents

About the Author

Andrew J. Grotto is the President and CEO of Sagewood Global Strategies LLC, a technology policy and risk advisory firm. He is also the founding director of the Program on Geopolitics, Technology, and Governance at Stanford University's Cyber Policy Center, a William J. Perry International Security Fellow, and a visiting fellow at the Hoover Institution. He serves as the faculty lead for the Cyber Policy and Security specialization within Stanford's Ford Dorsey Master's in International Policy Program and teaches the core course, Fundamentals of Cybersecurity and Policy. Before coming to Stanford in 2017, Grotto worked in public service, most recently as the Senior Director for Cyber Policy on the National Security Council at the White House.

grotto at sagewoodglobal.com
https://www.linkedin.com/in/andrew-grotto-2534b510a/
@GrottoAndrew

Author's Note and Disclosure

This working memorandum was commissioned by Workday to inform its views on the challenges and opportunities presented by growing policymaker and stakeholder interest in audit of AI/ML systems. The views contained in this memorandum, however, solely reflect the views of the author.

The author submits this to the FTC with the consent of Workday in the hopes that the Commission and the broader stakeholder community find it a useful contribution to the Commission's record.

I. Executive Summary

As societies grapple with how to govern artificial intelligence and machine learning technologies ("AI systems"), many stakeholders are looking to audit as a tool for guarding against bias, data protection, and other risks.

- **The New York City Council** passed a law in November 2021 requiring that employers conduct an independent audit of the automated tools they use to make employment decisions. The law goes into effect in 2023.
- **The Washington, DC City Council** is considering legislation loosely modeled on the New York law.
- **The California Civil Rights Department** is updating its fair employment and housing regulations to ensure that they cover decisions made by algorithms and to require employment decision makers to keep records.
- **Legislation at the U.S. federal level** introduced by Congressional Democrats in May 2022 on the use of automated decision tools by platforms includes a requirement that platforms keep records for potential review by the Federal Trade Commission (FTC) that document how it developed, tested and used the tools. In addition, the proposed bipartisan American Data Privacy and Protection Act includes requirements for covered entities and services providers to conduct an "algorithm design evaluation" (and, in earlier versions, a third party audit requirement).
- **The federal Equal Employment Opportunity Commission** (EEOC) issued technical guidance in May 2022 on the use of algorithms in the context of disability protections and is working on broader guidance; the guidance is likely to inform how other jurisdictions (e.g., New York City) craft audit criteria.
- **Internationally**, the European Commission's proposed Artificial Intelligence Act includes auditability requirements, while Europe's General Data Protection Regulation–which includes

protections for individuals against unwanted automated decision-making–has given rise to a vast compliance audit ecosystem.

An audit evaluates performance against expectations. It is a tool used in many sectors to advance various business and policy objectives.

- **The result of an audit is often a report** that describes a person or organization's conformance with established expectations. In some cases, audit findings may also result in a rating, seal, certification, or other trust mark that is intended to carry weight in the marketplace.

AI systems are information technologies and will often be embedded in digital systems that include AI and classical computing components.

- **An audit ecosystem for information technologies** already exists and does not generally distinguish AI systems from classical computing systems.
- **Too narrow of a focus on AI systems** risks missing interdependencies with classical computing components and other risk factors that affect the risk profile of a digital system, such as management and quality assurance controls, vulnerability to fraud perpetrated by trusted insiders or executives, supply chain risk management, and physical security.

Notes on Terminology

We use the phrases **"digital technologies"** and **"digital systems"** interchangeably to denote both information technologies (IT) and operational technologies (OT).

We use the phrase **"digital risks"** to refer to risks affecting digital technologies or systems.

An **"AI system"** is "an engineered or machine-based system that can, for a given set of human-defined objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy."

There is a robust research literature on audit: what works, what doesn't, and audit's limitations. Students of this literature have also identified various design considerations for an AI audit ecosystem.

Bottom line: The AI audit ecosystem is immature, at best. This conclusion is evident after reviewing the status quo audit ecosystem—especially the incumbent ecosystem for IT audit—against the design features that research has suggested are most associated with positive risk outcomes.

- **Technical and other challenges** are obstacles to generating audit results that relate to the harms people care about—overcoming them is an active area of research.
- **Audit criteria** for digital systems exist but how to apply them to AI systems is a work in progress.
- **Without auditable criteria** that map to settled expectations about AI system performance, an audit cannot purport to measure the system's performance against expectations.
- **The training, accreditation, and professional ethics ecosystem** for AI audit is in its infancy.

II. The Role of Audit

Audit is used in numerous industries and contexts to assess whether a claim of conformance to articulated specifications conforms to those specifications.

- **This usage of the phrase** is our focus here.[1]

Audit (as used in this memo) attempts to solve for a class of problems that emerge when a stakeholder has an interest in the actions of another party but has difficulties directly observing those actions and their effects. The stakeholder could rely on self-disclosures by the other party, but this is potentially risky because the other party, for reasons of incompetence, malice, or bias, may not be the most reliable source of information about their actions.

- **This is known as** an information asymmetry.
- **"Principal"** is a convenient shorthand for the beneficiary recipient of or audience for the audit, i.e the stakeholder.
- **"Agent"** is a convenient shorthand for the target of the audit, i.e. the actor whose decisions and actions the principal wishes to assess.

Information asymmetries emerge all the time in the economy. For example:

- **Senior managers** (the principal) may be at an information disadvantage vis-a-vis the various business units (the agent) under their supervision, if their only source of information about

[1] Social scientists may use the term "audit" differently. Danaë Metaxa and co-authors describe it as a social science testing methodology that involves "randomized controlled experiments in a field setting," in their recent overview of how the social science methodology could be applied to algorithms. Algorithm audit, according to this definition, involves "prob[ing] a process…by providing it with one or more inputs, while changing some attributes of that input," and studying the results. (Hackers might recognize this as a type of fuzzing.) Policy makers and researchers should note this definitional difference and strive for clarity about what they mean by "audit."

the performance of the business units is from the business unit itself.

- **Investors** (principal) may be at an information disadvantage vis-a-vis a firm's senior managers (agent), if their only source of information about the firm's performance is from those same senior managers.
- **Customers** (principal) may be at an information disadvantage when it comes to assessing the quality of a firm's (agent) products, if gauging quality is difficult without the sort of inside information that the firm supplying the products might have.[2]
- **Business partners** (principals) may be at an information disadvantage when it comes to evaluating each other's (agents) commitments to fulfilling their ends of a contract, especially in cases where a breach is difficult to detect before it causes harm. Much of the demand for audit of IT systems is driven by this problem.
- **Regulators** (principal) may be at an information disadvantage when it comes to preventing the harms that their rules are intended to guard against, if information about risk is primarily in the hands of the entities under their jurisdiction (agent).
- **The broader public** (principal) may be at an information disadvantage when it comes to assessing or verifying claims made by a public or private actor (agent) about a matter of public policy interest.
- **Sometimes, an actor is simultaneously** a principal to some agent (e.g., senior management as a principal to the business units within their span of responsibility) and an agent to some principal (e.g., senior management as an agent of the business' investors).

[2] If customers are unable to discern a desirable quality among competing products, they won't pay a premium for it and that quality will eventually disappear from the marketplace—what Nobel Prize-winning economist George Akerlof described as a "market for lemons."

Consider New York City's audit law: it requires employers to subject an automated employment decision tool to "a bias audit conducted no more than one year prior to the use of such tool" and to publicly disclose the results of the audit. Employers must also disclose to prospective job applicants how they use the tool.

- **These requirements are vaguely specified** in statute and are now the subject of rule-making.
- **The primary principals** for the law's audit requirements are job applicants in New York City. The agents are employers who use automated tools.
- **Most employers who use automated tools**, however, will have purchased them from a vendor, as opposed to developing them in-house. They will seek assurances from the vendors whose tools they have purchased that the tools can comply with (or enable the employer to comply with) whatever rules the New York City government ultimately promulgates; in this scenario, the employer is the principal and the vendor is the agent.
- **The complexity of this relationship matrix** may be compounded if the vendor uses data from the principal to train the tool: the employer is now part of the vendor's supply chain. The vendor may seek assurances from the employer about the quality of the training data, especially if use of the data exposes the vendor to any risk. If so, the vendor is a principal and the employer an agent when it comes to information asymmetries about the quality of the employer's training data.
- **Another layer of possible complexity:** Vendors may not have visibility into how an employer uses a software tool—this is typical for enterprise cloud software service providers and their customers. Vendors may seek assurances and protections against being held liable for how their customers use their tools. And even with legal protections in place, the vendor may still be vulnerable to reputational harm if one of their products is misused by a customer or inadvertently causes injury to third parties.

- **The broader public as a principal too?** The public has an interest in avoiding employment discrimination, and the law's requirement that the employer publish audit results is an invitation for public scrutiny of employers' use of algorithms and possibly litigation.

Bottom line: Audit has the potential to mitigate an information asymmetry by giving the recipient of the audit—the principal—the benefit of fuller information about whether the target of the audit—the agent—is living up to expectations.

- **Knowing who** the relevant principals and agents are, and what information asymmetries exist between them, is the starting point for designing any audit regime.

III. Audit and Digital Technologies: A Brief Overview

AI systems are digital technologies, for which there is already a robust, albeit imperfect, audit ecosystem.

- **The summary below** introduces the reader to this ecosystem, especially as it relates to the United States. It is not intended to be a comprehensive overview.

In 1969, the Electronic Data Processing Auditors Association (EDPAA) was founded in Los Angeles, CA. EDPAA would go on to become ISACA, which pioneered the field of civilian IT audit in the United States. It published the first version of "Control Objectives," an audit guide, in 1975 and introduced the Certified Information System Auditor (CISA) credential program in 1978.

The U.S. Department of Defense published the first of several iterations of "Trusted Computer System Evaluation Criteria"—more commonly known as the "Orange Book"—in 1983.

- **The Orange Book** specified standards for the federal government to assess a computer's security, which it used to make decisions about which computers could hold classified or other sensitive information, writes Steven B. Lipner in a short history of the initiative.
- **It defined three objectives:** give users a yardstick to assess how much trust could be placed in a computer that would store "classified or other sensitive information;" supply guidance to manufacturers as to how to satisfy classified use requirements; and provide those working on government acquisition with specific security requirements to inform their procurement contract negotiations.
- **It then laid out numerous controls** to guide assessments of computer security—against which companies could be audited.
- **The National Security Agency** performed the audits.

In the early 2000s, the U.S. government worked with counterparts in Canada, France, Germany, the Netherlands, and the United Kingdom to develop a new set of security evaluation standards, called the Common Criteria.

- **The Common Criteria** drew on the Orange Book and similar efforts by the other countries, specifically Canada's CTCPEC standard and the European ITSEC standard, as explained by Nancy Mead, a fellow at Carnegie Mellon University's Software Engineering Institute.
- **Unlike the DOD's cybersecurity audit guidelines** from the 1980s and 90s, the Common Criteria became an international standard through the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC)—specifically, ISO/IEC 15408.
- **The government's role in the audit process changed** as well. Instead of government agencies performing audits, a global network of mostly private testing laboratories does, with government agencies participating in only the most sensitive (and costly) reviews. The testing labs are accredited by national governments, which can also issue certificates for evaluated products that meet specified criteria; participating countries are supposed to recognize the validity of each others' certificates, regardless of which one of them issued it (this is referred to as mutual recognition).
- **Common Criteria is still in use**, having been updated most recently in August 2022.

Several major audit regimes are in use today for cybersecurity, data protection, asset management, quality assurance, and other risks affecting digital systems. Some of them are used by end-users of digital systems to demonstrate conformance with enterprise risk management standards; others are used by vendors of digital systems to demonstrate

that their products (or the processes used to develop them) conform to relevant product quality standards. [*See Appendix: Examples of Popular IT Risk Management and Audit Regimes*]

Various legal requirements have also given rise to audit regimes for digital systems, as actors subject to those laws seek assurances that they and/or their business partners are compliant.

- **Europe's General Data Protection Regulation** (GDPR), for example, has fueled a cottage industry of audit frameworks and auditors providing services to companies seeking to demonstrate conformance with the law.
- **In the U.S.,** laws that have given rise to audit ecosystems for digital systems include Sarbanes-Oxley for public companies; New York State's Department of Financial Services cybersecurity regulations; the Federal Financial Institutions Examination Council's cybersecurity regulations; and the security and privacy rules issued by the U.S. Department of Health and Human Services under the Health Information Portability and Accountability Act, to name just a few.

Bottom line: Audit for digital systems is not a new phenomenon, and indeed the audit ecosystem for digital technologies is already rich with standards, guidelines, and best practices, and also has a large, global community of auditors providing audit services against those expectations.

- **Key point:** This ecosystem does not generally distinguish digital technologies built with classical computing components from those built with AI components. As far as this ecosystem is concerned, AI systems are, at least in principle, already covered by it.
- **The efficacy of this incumbent ecosystem** at reducing digital risks**,** as we discuss below, is an open question.

IV. Ideal Features of an Audit Ecosystem

There is a robust research literature on audit: what works, what doesn't, and audit's limitations.

- **While few audit ecosystems** are composed wholly of these ideal design features, ecosystems regarded as healthy often have many of them.

Several ideal design features recur in this research literature:

1. **Audit results should relate** in a meaningful way to the risks that stakeholders actually care about. Otherwise, the audit is a compliance drill divorced from actual risk outcomes.
2. **Relatedly, the form that audit results take** should be intelligible to relevant stakeholders. (If not, what's the point of the audit?)
3. **The audit criteria should be defined with precision** and ideally have a binary, "yes or no" answer to whether a target conforms or not.
4. **The conditions that trigger an audit** should reflect the incentives and capabilities of principals, agents, and (especially) auditors. All parties incur direct and opportunity costs—principals must expend time and energy interpreting audit results; agents must expend time and energy dealing with the audit itself; auditors must expend time and energy to carry out the audit; and someone has to pay the auditor.
5. **The audit results should generally have a finite "shelf life"** of validity because circumstances that affect conformance can change over time: what may have been true during the audit could cease being true once the audit is complete.
6. **Auditors should have professional qualifications** that establish a floor for a minimum level of competence. There is a clear linkage between audit quality and the training, experience, and overall professionalization of auditors.

7. **Conflicts of interest** for auditors must be surfaced and managed, and will vary depending on whether the auditor is an independent, third-party with no interest in the cost, timeliness, or outcome of the audit or a non-independent second- or first-party auditor with some colorable interest in the cost, timeliness, or outcome of the audit that potentially gives rise to a conflict.

Go deeper: *Keep reading for a deeper dive into these design features for audit generally, which reflects a synthesis of previous literature reviews from Raji et al, Anderson (esp. chapters 12 and 28), Haapamäki and Sihvonen, and Slapinčar et al.*

- ***Or skip ahead*** *to Section V, "Audit and Digital Risks: A Closer Look," p.20.*

Going deeper…

1. Audit results should relate in a meaningful way to the risks that stakeholders actually care about.

- **In some cases,** the principal cares primarily about the agent's conformance with a set of standards because rote conformance is necessary to guard against reputational, contractual, or regulatory risks; whether conformance results in a reduction of other risks (e.g., to safety or security) may be immaterial to the principal.
- **In other cases,** a principal may believe that conformance is indeed linked to those other risk outcomes.
- **Another example:** consider an audit that examines whether an organization is implementing cybersecurity standards relating to the organization's resilience against cyber risks. Positive results may reassure senior management that the firm's information technology organization is meeting those standards, but the organization's resilience may not help a customer decide whether to trust the firm's assertions about the cybersecurity qualities of

its products: a company could be resilient against cyber threats, but still produce products that have cybersecurity problems.

- **If there is a mismatch** between a principal's belief about what an audit result implies about risk and what the result actually means, the audit results will be misleading.
- **Clarity about who the principal(s) are** and their information needs is therefore fundamental to the success of audit.

2. Relatedly, the form that audit results take should be intelligible to relevant stakeholders.

- **For example, an audit report** written in complex jargon is unlikely to meet a lay principal's information needs.
- **Efforts to translate complex audit results** into more intelligible forms often make use of seals, logos, certificates, and other trust markers.

3. The audit criteria should be defined with precision and ideally have a binary, "yes or no" answer to whether a target conforms or not.

- **Audits evaluate performance against expectations.** Those expectations, therefore, must be defined and be capable of being measured.
- **Measurement requires information,** and so auditors must have access to the information necessary to complete their audits. Agents may have incentives to share favorable information with the auditor but hide dispositive information.
- **Some critics have** challenged the applicability of the binary, "yes or no" norm to digital systems.

4. The conditions that trigger an audit should reflect the incentives and capabilities of principals, agents, and auditors.

- **These conditions can take a variety of forms.** An actor may seek an audit voluntarily, in response to market forces, or be required to submit to one under the law. In cases where audit is a legal

requirement, it could apply to an entire category of actors, be triggered by an adverse event, be conducted randomly, or pursued according to a risk-informed algorithm.

- **Direct and opportunity costs are practical constraints** on audit. Principals must expend time and energy interpreting audit results; agents must expend time and energy dealing with the audit itself; auditors must expend time and energy to carry out the audit; and someone has to pay the auditor.
- **For regulators,** the methodology for deciding how to select targets for audit must be sensitive to the regulator's institutional capacity for carrying out or supervising audits.

5. The audit results should generally have a finite "shelf life" of validity because circumstances that affect conformance can change over time: what may have been true during the audit could cease being true once the audit is complete.

- **The length** of the term should be linked to how rapidly circumstances change and the effects of those changes on the reasonable expectations of stakeholders in the audit results. Considerations of audit costs may also be relevant—audits cost money and can be a drag on other business objectives.

6. Auditors should have professional qualifications that establish a floor for a minimum level of competence. There is a clear linkage between audit quality and the training, experience, and overall professionalization of auditors.

- **In some audit domains,** such as financial statement auditing, auditors may be required to have met certain accreditation or licensing requirements.
- **Auditor education curricula and credentials** exist for various audit domains, through government regulators (e.g., state boards of accountancy), professional associations (e.g., Association of International Certified Professional Accountants), and

international standards bodies (e.g., International Accounting Education Standards Board).

- **Standards relating to audit quality** also exist in numerous forms. There are standards for different audit methodologies (e.g., for internal audit, the Risk Management Framework from the Committee of Sponsoring Organizations); standards for different audit targets (e.g., for quality management processes, the ISO 9000 series); standards for managing certain risks (e.g., for information security management, ISO/IEC 27001); standards for auditing specific products (e.g., for automotive quality management, IATF 16949); and standards applicable to specific sectors (e.g., for public companies, the IPSASB International Public Sector Accounting Standards).
- **Laws and regulations** relating to audits often incorporate standards by reference and may also set forth requirements for how organizations, as part of their effort to comply with the law, carry out audits.
- **Supervision and enforcement** mechanisms of audit standards exist for some audit domains, most notably for public companies. In the United States, for example, the Public Company Accounting Oversight Board is responsible for supervising the audits of public companies and SEC-registered brokers and dealers. Pursuant to this mission, it sets audit standards, performs inspections of audit firms, and disciplines auditors who commit malpractice.

7. **Conflicts of interest** for auditors must be surfaced and managed.
   - **Who performs the audit** and that auditor's incentive structure matter because conflicts of interest can emerge that potentially influence the auditor's behavior.
   - **Professional, enforceable standards of conduct** can help mitigate conflicts by establishing what is appropriate behavior and holding auditors who violate those standards accountable. Organizations such as the International Ethics Standards Board for Accountants

and the American Institute of Certified Public Accountants maintain ethical guidelines for auditor independence.

7.a Independent, third-party audit (I3A) is the gold standard for auditing from a conflicts of interest perspective, but it is used comparatively rarely. It involves an external auditor with no interest in the cost, timeliness, or outcome of the audit.

- **Their sole performance incentive** is to uphold professional standards of conduct for conducting audits. Their ability to attract and retain audit engagements should not otherwise depend on pleasing the principal or the agent.
- **These are demanding conditions.**

7.b Second-party audit (2PA) is far more common. It too involves an external third-party auditor, except that the auditor is not independent—they have some colorable interest in the cost, timeliness, or outcome of the audit that potentially gives rise to a conflict. Conflicts and the resulting bias can emerge in subtle and not-so-subtle ways for 2PA. For example:

- **Who's paying** the audit bill? If it's the agent, the auditor faces an obvious financial conflict of interest between getting paid for the engagement and delivering audit results that jeopardize their business relationship. A less obvious but documented conflict can also emerge when it's the principal paying the auditor, if the auditor is under any pressure to deliver results that confirm a principal's prior claims, expectations, or beliefs.
- **Do auditors compete** for business? If so, there is a well-documented risk that competition can fuel a "race to the bottom" where auditors, in a quest for repeat business, compete on price, timeliness, and ease of passage as opposed to audit quality.
- **Are audit services bundled** with other services, such as consulting? If so, the auditor's quest for consulting business can also contribute to a race to the bottom for audit quality.

7c. First-party audit (1PA) is also common; it is where an organization uses an internal auditor.

- **When evaluating the risk of conflicts** for 1PA, knowing who the relevant principal(s) are is especially important. An internal audit conducted for internal principals who bear the full cost of their actions—including poor choice or oversight of internal auditors—may be less risky from a conflicts perspective than cases where some or all of those costs are borne by third parties who cannot affect the performance of the auditors.

Bottom line: Clarity about the interests and capabilities of different principals, their agents, and auditors is essential to designing an audit regime that empowers principals to make better, more informed decisions—and hold agents who fall short of expectations accountable.

V. Audit and Digital Risks: A Closer Look

The main driver of interest in AI audit among policymakers and other concerned stakeholders is its potential to mitigate harms that digital systems with AI components could cause to people and societies.

- **In other words,** audit is a tool and audit results should relate positively to better risk outcomes.

Passing an audit does not necessarily mean the target is more resilient against digital risks.

- **Many organizations have received positive audit results** that report conformance with leading digital risk standards and best practices, but still suffered major incidents involving security shortcomings that a qualified auditor arguably should have identified.
- **Equifax and Target are two prominent examples.** EY performed an audit of Equifax against ISO 27001 and deemed it compliant; Equifax subsequently suffered a breach that compliance with ISO 27001 arguably should have prevented. Target suffered a major breach of credit card information in November 2013, even though (in the words of Target's then-CEO) "Target was certified as meeting the [security] standard for the payment card industry (PCI) in September 2013," two months prior to the breach.

The various explanations for how incidents such as these could happen highlight the challenges of auditing digital systems.

- **The auditor might have missed problems.** Yet there is no public information about auditors suffering any meaningful consequences that we are aware of. (Students of financial auditing for public companies have raised similar concerns about enforcement of professional standards for financial auditors.) ForHumanity, a non-profit organization in the early stages of developing an audit ecosystem for AI, has issued professional

standards of conduct for AI audit, but monitoring and enforcing compliance with those standards will be a challenge.

- **Circumstances might have changed** in the period between when the audit results came in and the incident happened. If so, the shelf life of the audit was arguably too long. No and Vasarhelyi document the many challenges that digital systems present to the traditional, point-in-time audit report. The trend in cyber risk management is towards continuous monitoring, with companies such as SecurityScorecard and BitSight offering continuous monitoring services in the form of security ratings.
- **Or, the audit criteria do not fully capture the risk** exposure of the audit target. This would suggest that the audit results do not correspond to the risks that stakeholders care about and/or the audit criteria have gaps.

Overall, there is limited empirical research on the relationship between audits of digital systems and risk outcomes.

Most studies focus on first-party or second-party audit (1PA and 2PA, respectively–see "Going Deeper" in Section IV for further background) for cybersecurity risk.

- **In their literature review of cybersecurity in accounting** and audit research from 2019, Haapamäki and Sihvonen identified 39 high-quality studies; nearly all of them involve 1PA or 2PA aimed at supporting a firm's management decisions about risk. They conclude that "research on the role of cybersecurity in private and public companies is still relatively scarce." They also acknowledge the limitations of existing research when they call for "future studies [that] investigate how the validation of the disclosed information is performed and what role auditors perform in cybersecurity risk management…[and] how the training and competence of auditors related to cybersecurity might be improved."

- **Another recent (2022) study developed an index** to quantify the use of cybersecurity audits and found no statistical relationship between the use of cybersecurity audits and the probability of suffering a cybersecurity incident—a discouraging result. On the other hand, the study does not purport to offer insights into a relationship between audit and the severity of the consequences of a cybersecurity incident—an important caveat.

Adverse selection, where the actors who seek positive audit results do so in order to obscure or legitimize problematic behavior, is a documented challenge for digital technologies.

- **For example, an earlier study found** that websites that had earned the TRUSTe certification of site integrity were more likely to spread malware to visitors. The problem in this case appears to have stemmed from the lax qualification criteria for receiving the certification and the purveyors of the certification's lack of oversight. In essence, the lack of rigor attracted rogues who exploited the positive connotations of the certification to perpetrate various cyber crimes.
- **More recently,** an episode involving the AI company HireVue highlights how information asymmetries, vague audit criteria, and use of a 2PA can result in misleading audit results. HireVue paid a firm to audit its flagship product, an algorithm marketed as helping employers make hiring decisions; the product incorporated controversial facial recognition technologies. HireVue allegedly misrepresented the results of the audit, which itself was based on a curated information set provided by HireVue.
- **Adverse selection and its cousin, moral hazard**, are major concerns for the cyber insurance industry, especially in the context of ransomware attacks. For example, one concern is that an organization with insurance that covers ransomware payments may be tempted to skimp on costly efforts to prevent ransomware

attacks if the organization believes that an insurance company will foot the bill for an attack.

The Orange Book's shortcomings became evident with time and experience.

- **Some of the problems concerned audit criteria.** For example, there was an issue of "criteria creep" where new computer systems demanded new security requirements, and soon the criteria needed to be expanded and/or interpreted widely—making the guidance far more complex. The audit criteria were also tied to the technology of the day, focusing mainly on operating system security issues; when risks outside the operating system grew in the 1990s, the original criteria became less relevant (or, at least, less comprehensive).
- **Other problems involved the audit process,** as Steven Lipner documents in his aforementioned history of the Orange Book. The Orange Book emphasized mandatory security controls and a high degree of assurance—that is, ensuring that a computer has sufficient hardware and software to be properly evaluated. However, this made it time-consuming and costly for manufacturers to get evaluated, and many companies gradually lost interest in passing Orange Book audits as the global market for their products expanded and the U.S. government's relative degree of purchasing power diminished.
- **Ross Anderson** (in chapter 28) aptly summarizes these problems:

> "When the Orange Book was written, the Department of Defense thought that they paid high prices for high-assurance computers because the markets were too small, and hoped that security standards would expand the market. But Orange Book evaluations followed government work practices. A government user would want some product evaluated; the NSA would allocate people to do it; given

> traditional civil service caution and delay, this could take two or three years; the product, if successful, would join the evaluated products list; and the bill was picked up by the taxpayer. Evaluated products were always obsolete, so the market stayed small, and prices stayed high."

Lipner's grim conclusion about the Orange Book: "If the objective […] was to create a rich supply of high assurance systems that incorporated mandatory security controls, it is hard to find that the result was anything but failure."

Common Criteria has come under fire for similar reasons, with the added problem of testing laboratories—the auditors—competing with each other on price and timeliness, which fuels concern about a "race to the bottom" for audit quality.

- **Ross Anderson pulls no punches** in his extended critique (chapter 28) of Common Criteria, concluding that "the operation of CC outside Europe has been a bit of a joke, and even within Europe it has been undermined by both companies and countries gaming the system."

IT auditors face many challenges as well. Emerging technology and infrastructural changes can make audits complicated, and cybersecurity threats add a dynamic, adversarial dimension, according to an industry survey. Auditors without the requisite staffing and skills face hurdles, and auditors on top of that may have to manage tricky relationships with third parties and vendors touching the company's IT systems.

- **For example, Barclay Simpson found** in a 2016 survey that internal auditors face increasing workload demands, and many internal audit departments are inadequately resourced. Other surveys reach similar conclusions.

Accreditation programs for systems administration, cybersecurity, and privacy/data protection are popular and valued in the marketplace.

- **Over 90% of IT professionals** in a recent survey by Skillsoft report having at least one professional certification.
- **Most professional certifications for IT** involve risk management or skills such as cloud services administration or penetration testing. Audit is a more niche field, though the Certified Information Systems Auditor (CISA) credential ranks among the most popular certifications among the IT professionals surveyed by Skillsoft.
- **AI audit could piggyback on these programs,** especially for cybersecurity and other digital risks for which there is already a relatively mature standards and audit ecosystem.
- **Research is underway** on how to adapt existing IT risk management policies and resources to AI systems, but the field is in its early stages.

Bottom line: An audit ecosystem already exists for digital technologies, with cybersecurity risk standing out as having the densest and arguably the most mature collection of standards, guidelines, best practices, and accreditation architecture.

- **Even so,** publicly available evidence linking IT audit outcomes to IT risk outcomes is ambiguous, let alone whether one set of standards is associated with better risk outcomes than another.
- **To be sure,** this does not necessarily mean that IT audit outcomes are divorced from risk outcomes: it is entirely possible, and indeed quite plausible, that organizations seeking successful audits against standards produced by reputable organizations have better risk outcomes than those who do not. This could reflect, however, management's focus on managing risk as opposed to rote compliance with this or that standard.
- **Audit frameworks can also be gamed**, as the examples described above warn.

VI. Conclusions

Audit of AI systems has the potential to resolve information asymmetries that frustrate the ability of stakeholders to exercise appropriate oversight of AI systems.

- **Much work, however,** remains to be done. When measured against the ideal design features that students of audit have identified, the AI audit ecosystem is immature, at best.
- **Moreover, even though** its natural overarching ecosystem—the digital technologies audit ecosystem—is relatively mature, in terms of the density of auditable standards, guidelines and best practices, the empirical evidence for audit yielding better risk outcomes is limited and mixed.

Bottom line: The AI audit ecosystem is immature, at best.

- **Technical and other challenges** are obstacles to generating audit results that relate to the harms people care about.
- **Audit criteria** for digital systems exist but how to apply them to AI systems is a work in progress. ForHumanity has developed a UK GDPR Certification Scheme that includes AI systems as audit targets as well as a UK Children's Code Certification Scheme. Researchers have also proposed various audit frameworks for ethical risks such as bias and discrimination, but there is no consensus on them yet, let alone on criteria for other risks that AI systems could give rise to, such as data protection, cybersecurity, physical injury, and other harms.
- **Without auditable criteria** that map to settled expectations about AI system performance, an audit cannot purport to measure the system's performance against expectations.
- **The training, accreditation, and professional ethics ecosystem** for AI audit is in its infancy.

At worst, unrealistic expectations from policymakers and the public about the current potential for audit as a tool for reducing AI risk could result in misplaced confidence in audit results and undermine the credibility of audit generally.

Appendix: Examples of Popular IT Risk Management and Audit Regimes

COBIT, or the Control Objectives for Information and Related Technologies, was published in 1996 by ISACA.

ISO's 27000 series grew out of information security standards published by the British Standards Institute in 1995. The standards in the 27000 "series," as ISO calls a grouping of standards, provide a framework for information security on IT systems.

Service Organization Control (SOC) reports are prepared in accordance with standards developed by the American Institute of Certified Public Accountants (AICPA) relating to internal controls. Initially focused on financial reporting, AICPA has adapted the SOC methodology to cybersecurity and supply chain risk management. SOC-1 reports focus on a company's internal controls around financial statements; SOC-2 reports focus on internal controls related to data processing, including data security, confidentiality, and privacy; and SOC-3 reports focus on conveying SOC-2 results to a general audience.

SOC-2 reporting is popular for cybersecurity. It is based on standards by the Auditing Standards Board of the American Institute of Certified Public Accountants (AICPA), and it is intended to align with popular standards used to generate attestations about an organization's controls (specifically, AICPA's SSAE 18 and the IAASB's ISAE 3402). For example, some companies use SOC-2 reports to audit their cloud security posture, finding value in a SOC-2 approach because it allows the auditors to communicate information on controls to current and prospective customers, develop risk mitigations, and adapt to customer interests and needs. The AICPA also advertises SOC reports as helping senior management, boards of directors, and investors, among others, to understand cybersecurity controls within an organization.

NIST, the U.S. National Institute of Standards and Technology, has developed and maintains a catalog of publications that specify risk controls and methodologies for federal government and federal government contractor digital systems. Some private organizations implement NIST requirements even if they have no federal contracts because NIST's publications carry some weight in the marketplace. (NIST also develops risk management frameworks for government and private sector use—such as the Cybersecurity Framework and the draft AI Risk Management Framework—but they are not designed to be audited against.)

ENISA, the European Union's cybersecurity agency, is charged under EU law with developing certification frameworks for digital systems. It transmitted the first completed framework—an implementation of the Common Criteria—to the European Commission in May 2021 after nearly two years of development effort.

Bibliography

1. American Institute of Certified Public Accountants (n.d.), "SOC for Service Organizations," https://us.aicpa.org/interestareas/frc/assuranceadvisoryservices/socforserviceorganizations.html.
2. American Institute of Certified Public Accountants (n.d.), "Clarified Statements on Standards for Attestation Engagements," https://us.aicpa.org/research/standards/auditattest/ssae).
3. J. P. Anderson (1972), "Computer Security Technology Planning Study (Volume II)," https://csrc.nist.gov/csrc/media/publications/conference-paper/1998/10/08/proceedings-of-the-21st-nissc-1998/documents/early-cs-papers/ande72.pdf).
4. R. Anderson (1993), "Why Cryptosystems Fail," in *Proceedings of the 1st ACM Conference on Computer and Communications Security*, *CCS '93*, pp. 215–227.
5. R. Anderson (2001), "Why Information Security Is Hard - An Economic Perspective," in *Proceedings of the 17th Annual Computer Security Applications Conference,* pp. 358–365, https://www.cl.cam.ac.uk/~rja14/Papers/econ.pdf.
6. R. Anderson (2020), *Security Engineering: A Guide to Building Dependable Distributed Systems* (Wiley, 3rd Edition).
7. R. Anderson, T. Moore (2006), "The Economics of Information Security," https://tylermoore.utulsa.edu/science-econ.pdf.
8. H. Asghari, M. J. G. van Eeten, J. M. Bauer (2016), "The Economics of Cybersecurity," in *Handbook on the Economics of the Internet*, J. M. Bauer, M. Latzer, eds. (Edward Elgar Publishing, Cheltenham and Northhampton), pp. 11–40, https://hadiasghari.com/mirror/asghari16_econsec.pdf,
9. Barclay Simpson (2016), "Internal Audit Market Report 2016," http://www.barclaysimpson.com/File.ashx?path=Root/Documents/IntAudit2016_fin.pdf.

10. S. Brown, J. Davidovic, A. Hasan (2021), "The algorithm audit: Scoring the algorithms that score us." *Big Data & Society* 8, doi:10.1177/2053951720983865.
11. R. Calo (2017), "Artificial Intelligence Policy: A Primer and Roadmap." *UC-Davis Law Review* 51, 399.
12. A. Chaudhary (2020), "Using SOC Reports for Cloud Security and Privacy," *Industry Insights*, https://cloudsecurityalliance.org/blog/2020/02/10/using-soc-reports-for-cloud-security-and-privacy/.
13. Common Criteria (n.d.), "The Common Criteria," available at https://www.commoncriteriaportal.org/.
14. A. Daly *et al* (2019)., "Artificial Intelligence Governance and Ethics: Global Perspectives," arXiv:1907.03848.
15. D. P. David, A. Mermoud, S. Gillard (2021), "Cyber-Security Investment in the Context of Disruptive Technologies: Extension of the Gordon-Loeb Model," 1–18, http://arxiv.org/abs/2112.04310.
16. J. X. Dempsey, A. J. Grotto (2021), "Vulnerability Disclosure and Management for AI/ML Systems: A Working Paper with Policy Recommendations", https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/ai_vuln_disclosure_nov11final-pdf_1.pdf.
17. S. Drimer, S. J. Murdoch, R. Anderson (2008), "Thinking Inside the Box: System-Level Failures of Tamper Proofing," in *2008 IEEE Symposium on Security and Privacy*, pp. 281–295.
18. B. Edelman (2011), "Adverse Selection in Online "Trust" Certifications and Search Results," *Electronic Commerce Research and Applications* 10, 17–25 (2011).
19. A. Engler (2021), "Auditing employment algorithms for discrimination," https://www.brookings.edu/research/auditing-employment-algorithms-for-discrimination/.
20. ENISA (2021), "Cybersecurity Certification: Candidate EUCC Scheme V1.1.1 — ENISA,"

https://www.enisa.europa.eu/publications/cybersecurity-certification-eucc-candidate-scheme-v1-1.1.
21. ForHumanity Ethics Committee (2021), "ForHumanity Certified Auditors (FHCA) Code of Ethics," https://forhumanity.center.
22. Gartner (2022), "Gartner Survey Reveals the Top Challenges for Internal Audit in 2022," https://www.gartner.com/en/newsroom/press-releases/2022-03-17-gartner-survey-reveals-the-top-challenges-for-internal-audit-in-2022.
23. L. A. Gordon, M. P. Loeb (2002), "The Economics of Information Security Investment." *ACM Transactions on Information Systems Security* 5, 438–457.
24. A. A. Gramling, M. J. Maletta, A. Schneider, B. K. Church (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, 194–244.
25. A. Grotto, G. Falco, I. Maifeld-Carucci (2021), "Response to 'Request for Information: Artificial Intelligence Risk Management Framework' (86 FR 40810)," https://www.nist.gov/system/files/documents/2021/09/16/ai-rmf-rfi-0077.pdf.
26. E. Haapamäki, J. Sihvonen (2019), "Cybersecurity in accounting research." *Managerial Auditing Journal* 34, 808–834.
27. T. Hagendorff (2019), "The Ethics of AI Ethics: An Evaluation of Guidelines," https://arxiv.org/pdf/1903.03425.pdf.
28. D. S. Hilzenrath, N. Trevino (2019), "How an Agency You've Never Heard of Is Leaving the Economy at Risk," https://www.pogo.org/investigation/2019/09/how-an-agency-youve-never-heard-of-is-leaving-the-economy-at-risk.
29. International Auditing and Assurance Standards Board (2010), "ISAE 3402 Assurance Reports on Controls at Service Organizations."

30. ISACA (n.d.), "ISACA History & Timeline," https://www.isaca.org/why-isaca/about-us/isaca-50/timeline).
31. ISACA, Protiviti (2019), "2019 Audit Benchmarking Study."
32. ISACA, Protiviti (2022), "IT Audit Perspectives on Today's Top Technology Risks," https://www.protiviti.com/hk-en/survey/2022-global-finance-trends-survey.
33. S. B. Lipner (2015), "The Birth and Death of the Orange Book." *IEEE Annals of the History of Computing* 37, 19–31.
34. P. Lovaas, S. C. Wagner (2012), "IT Audit Challenges for Small and Medium-Sized Financial Institutions," *Annual Symposium on Information Assurance and Secure Knowledge Management* 16-22.
35. N. Mead (2006), "The Common Criteria," www.sei.cmu.edu.
36. D. Metaxa *et al.* (2021), "Auditing Algorithms: Understanding Algorithmic Systems from the Outside In." *Foundations and Trends in Human-Computer Interaction.* 14, 272–344.
37. D. G. Mihret, B. Grant (2017), "The Role of Internal Audit in Corporate Governance: A Foucauldian Analysis." *Accounting, Auditing & Accountability Journal* 30, 699–719 (2017).
38. T. Moore, R. Anderson (2012), in *The Oxford Handbook of the Digital Economy*, M. Peitz, J. Waldfogel, eds., pp. 572–599. https://tylermoore.ens.utulsa.edu/oxford12.pdf.
39. National Institute of Standards and Technology (n.d.), "Publications," https://csrc.nist.gov/publications).
40. National Security Commission on Artificial Intelligence (2021), "Final Report."
41. W. G. No, M. A. Vasarhelyi (2017), "Cybersecurity and Continuous Assurance." *Journal of Emerging Technologies in Accounting.* 14, 1–12.
42. OQRI (2018), "Equifax Held ISO 27001 Certification At Time of Massive System Hack - Oxebridge Quality Resources," https://www.oxebridge.com/emma/equifax-held-iso-27001-certification-at-time-of-massive-system-hack/.

43. ORCAA (2020), "Description of Algorithmic Audit: Pre-built Assessments," 1–8.
44. I. D. Raji *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).
45. I. D. Raji, P. Xu, C. Honigsberg, D. Ho (2022), in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AEIS'22)*, pp. 1–16.
46. M. Riley, B. Elgin, D. Lawrence, C. Matlack (2017), "Missed Alarms and 40 Million Stolen Credit Card Numbers: How Target Blew It." *Bloomberg*, https://www.bloomberg.com/news/articles/2014-03-13/target-missed-warnings-in-epic-hack-of-credit-card-data.
47. A. Sayana (2022), "The evolution of information assurance." *ISACA J.* 1, 1–5.
48. H. Schellmann (2021), "Auditors are testing hiring algorithms for bias, but there is no easy fix." *MIT Technology Review*, https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/.
49. H. R. K. Skeoch (2022), "Expanding the Gordon-Loeb model to cyber-insurance." *Computer Security* 112, 102533.
50. S. Slapničar, T. Vuko, M. Čular, M. Drašček (2022), "Effectiveness of cybersecurity audit." *International Journal of Accounting Information Systems* 44, 100548.
51. G. Stults (2004), "An Overview of Sarbanes-Oxley for the Information Security Professional," https://sansorg.egnyte.com/dl/XYaFkPDbB7.
52. P. Sullivan (2021), "The Most Common Challenges of the Audit Process," https://www.a-lign.com/articles/common-challenges-audit-process.

53. O. Turetken, S. Jethefer, B. Ozkan (2020), “Internal Audit Effectiveness: Operationalization and Influencing Factors.” *Managerial Auditing Journal* 35, 238–271.
54. United States Department of Defense (1985), “Trusted Computer System Evaluation Criteria ["Orange Book"].”