Public vs Private Bodies: Who Should Run Advanced AI Evaluations and Audits? A Three-Step Logic Based on Case Studies of High-Risk Industries

Merlin Stein[1*], Milan Gandhi[1], Theresa Kriecherbauer[1], Amin Oueslati[2] and Robert Trager[1]

[1]University of Oxford
[2]Hertie School
merlin.stein@bsg.ox.ac.uk

Abstract

Artificial Intelligence (AI) Safety Institutes and governments worldwide are deciding whether they evaluate and audit advanced AI themselves, support a private auditor ecosystem or do both.

Auditing regimes have been established in a wide range of industry contexts to monitor and evaluate firms' compliance with regulation. Auditing is a necessary governance tool to understand and manage the risks of a technology. This paper draws from nine such regimes to inform (i) who should audit which parts of advanced AI; and (ii) how much capacity public bodies may need to audit advanced AI effectively.

First, the effective responsibility distribution between public and private auditors depends heavily on specific industry and audit conditions. On the basis of advanced AI's risk profile, the sensitivity of information involved in the auditing process, and the high costs of verifying safety and benefit claims of AI Labs, we recommend that public bodies become directly involved in safety critical, especially gray- and white-box, AI model evaluations. Governance and security audits, which are well-established in other industry contexts, as well as black-box model evaluations, may be more efficiently provided by a private market of evaluators and auditors under public oversight.

Secondly, to effectively fulfill their role in advanced AI audits, public bodies need extensive access to models and facilities. Public bodies' capacity should scale with the industry's risk level, size and market concentration, potentially requiring 100s of employees for auditing in large jurisdictions like the EU or US, like in nuclear safety and life sciences.

1 Introduction

Governments across the world are exploring new regulations to mitigate the risks of advanced artificial intelligence (Weidinger et al. 2022, UK Government 2023). Establishing an auditing regime is one tool available to policymakers to facilitate and enforce firms' compliance with AI rules. For the purposes of this paper, we define an 'AI auditing regime' as the institutional framework by which advanced AI developers and providers ("AI Labs"), particularly their AI models, are subjected to evaluation by externals. Under this definition, there are numerous design choices available to policymakers (Birhane et al. 2024). We explore two sets of choices, (1) Who should audit which parts of advanced AI? (2) What resources, competence and access should the public body develop to carry out its role in auditing? We use 'public body' to refer to the government institution that is primarily responsible for auditing, whether through rule-setting and oversight or undertaking audits. *We use the term "audit" to include exploratory evaluations, targeted auditing and monitoring.* Our contributions are:

- **Auditing Regime Case Analysis and Design Factors:** We analyze nine industry cases to identify dimensions along which auditing regimes differ, and quantify industry and audit factors explaining the differences. These are an extension of hybrid governance theory.
- **Three-Step Logic for Auditing Regime Design:** We propose a three-step logic to determine who is best placed to audit depending on the industry context, demand for auditing and the type of auditing required. We apply this logic to derive policy recommendations for designing advanced AI auditing regimes.
- **Estimate of Required Capacity in AI Safety Institutes or Other Public Bodies for Advanced AI:** We empirically estimate the resource, competence and access requirements for a public body in an advanced AI auditing regime.

This paper is structured as follows:

- Section 2 locates the study in the literature
- Section 3 explains the methodology and limitations
- Section 4 proposes demand-side and supply-side factors determining who could and should audit
- Section 5 explores nine high-risk auditing regimes and extrapolates a three-step logic on who should audit
- Section 6 applies the three-step logic to advanced AI
- Section 7 outlines resource, competence and access requirements for public bodies in advanced AI auditing
- Section 8 describes open questions and concludes

*Primary author, majority contribution

2 Related Literature

We use the term 'advanced AI' to refer to state-of-the-art general-purpose AI models, aligning with the definition of the International Scientific Report on the Safety of Advanced AI (DSIT 2024). As this report and other research analyzes, firms that develop advanced AI risk the imposition of unpredictable and potentially severe costs on unconsenting third parties (DSIT 2024). Such externalities require government intervention (Pigou 1920). Embedded in a spectrum of measures (Gunningham, Grabosky and Sinclair 1998), one important intervention is AI auditing (Costanza-Chock et al. 2023). However, advanced AI is one of the fastest evolving and complex general-purpose technologies. Its externalities are difficult to reliably estimate (DSIT 2024, Hobbhahn and Scheurer 2024). Thus, advanced AI auditing regimes need to address the challenge of running the *right* audits *well* under resource constraints ("audit effectiveness")[2].

Running the right audits. Sufficient and flexible capacity is necessary to keep up with the speed of AI progress and therefore an expanding list of dangerous capabilities and downstream sociotechnical risks (EpochAI 2023). Auditors need to be competent and have access to assess capabilities and risks. Running the right audits means reducing uncertainties, e.g. through standardization.

Running audits well: Independence vs. resource efficiency. The most independent auditors aligned with public interest – public bodies, publicly-appointed auditors, academics or civil society – may be less efficient and flexible than private auditors. However, private auditors fail to produce high quality auditing when they share conflicts of interest with auditees (DeFond 2010). Thus, balancing independence and efficiency can mean trading-off audit quality and efficiency. This trade-off shapes auditing regimes. Audit quality and independence is more important for audit steps that are critical for public safety (Brundage et al. 2020, Power 1999). The industry setting and auditing ecosystem, including the distribution of resources and skills may influence this trade-off (Power 1999).

Previous literature observes significant variability in the design and effectiveness of auditing regimes across industry contexts (Kleinman, Lin, and Palmon 2014; Raji et al. 2022). Key factors influencing auditing effectiveness include auditor independence (Duflo et al. 2013, Short et al. 2016), resources (Anderljung et al. 2023), competence (DeFond and Zhang 2014) and the auditor's access to evidence for the audit (Lamoreaux 2016, Raji et al. 2022), and the auditor's access to the evidence required for audits (Simnett, Carson, and Vanstraelen 2016; Simnett and Trotman 2018; Hansen, Kumar, and Sullivan 2008).

Questions about the public body's optimal role in an advanced AI auditing regime remain under-explored (Hadfield and Clark 2023). Extant auditing literature emphasizes auditor characteristics like independence in explaining regime effectiveness. Our research examines underlying characteristics of the industry and audit, exploring their implications for auditing regime design.
For this purpose, we connect with the hybrid governance literature and new institutional economics. Effective governance is only partly determined by the characteristics of the oversight or auditing body, mainly by the alignment of these characteristics with the underlying conditions of transaction costs and asset specificity (Menard 2004, Quélin et al. 2019). As hybridity shapes AI governance (Radu 2021) and auditing too (Rajala and Kokko 2021), we adapt Menard's (2004) hybrid governance framework for the auditing context.

3 Methodology, Scope and Limitations

To analyze which advanced AI audits should be performed by public and which by private bodies and its resource implications, we surveyed examples of auditing regimes across nine different industries, focusing our analysis on critical infrastructure sectors in the United States ("US")[3]. This comparative case study approach has proven effective

[2] An effective audit requires accurately assessing relevant benefits and harms ("audit quality"), while minimizing costs and delay ("audit efficiency").

[3] We focus on US oversight regimes, given the country's leading AI development capacity (Tortoise 2023). As part of the case studies, we briefly compared each industry's US regimes to their counterparts in the EU and UK, finding no major deviations, even though regimes in the US are slightly more liberal in most industries (e.g., OECD 2014). We add audits of online platforms in the EU as an additional case that is less present in the US.

The US, for example, categorizes AI as a critical and strategically important technology (NSTC 2024). Given Advanced AI's potential risks, we focus on sectors classified as critical infrastructure in the US, EU and UK. Of these, we use sectors with especially high speed of innovation based on patents filed (Marco et al. 2017). Within each sector, we pick a typical product or security system for clarity. This leads to our case studies on transport (airworthiness certification of civil airplanes), communications (authorization of radio frequency devices in telecommunications), IT (cyber security infrastructure - for nuclear power plants, government contractors and bulk power systems; and large online platforms), financial services (audits of public companies' annual reports, risk monitoring of financial securities products), and life sciences (regulatory approvals process for medical software). We focus on audits in civilian contexts, within a single jurisdiction, aimed at general-purpose models. We assume that advanced AI will be developed by private entities.

for similar prescriptive questions on regulatory regimes (Levi-Faur 2003, Hill and Varone 2021). Given the small number of high-risk regimes and difficulty in capturing nuances in their variations quantitatively, we deploy an exploratory, inductive mixed-method approach. Based on existing literature, case studies and in line with hybrid governance theory, we identify demand-side factors determining auditing responsibilities across and within industries. To understand variations at a high level, we quantify the demand-side factors for each case, and observe their association with the degree of public body involvement in auditing. To explain this link and derive more granular implications for advanced AI auditing, we qualitatively analyze auditing supply and estimate public bodies' capacity requirements.

Our case study research describes what *is* the case across contexts, and does not measure effectiveness of audit regimes directly, nor establish a causal link quantitatively. Instead, we follow Menard (2004) and assume that effective governance is largely determined by the alignment of the characteristics of the auditing body with underlying demand-side factors. Further, our capacity estimates are initial, rough approximations, and require more dedicated research.[4] Appendix B further details the methodology and limitations.

4 Framework of Analysis: Auditing Factors

We propose that differences across regimes on *who audits* and *with which resources* can be understood by considering the nature of risks in an industry being addressed by the audit and challenges inherent to auditing (demand-side factors); and the characteristics of auditors (supply-side factors). We analyze them in each case study.

4.1 Demand-Side Factors: The Nature of the Risk

We suggest that the demand for and emergence of an auditing regime is shaped by an industry's risk profile and its perceived importance by the public (Ramanna 2015). In addition, the industry market size and concentration may impact the volume of audits demanded.

An adjacent set of demand-side factors are inherent to a specific audit. These relate to the availability of auditors and the skills they require to conduct audits, which vary according to the complexity of auditing methods. Furthermore, there is a perennial information problem – to conduct the audit, the auditor must obtain information in the control of the auditee and verify the auditee's claims. Finally, collecting, managing and, in some regimes, publishing this information may pose its own set of risks if the information is sensitive to, for example, intellectual property and national security concerns.

Demand-side factors	
Industry conditions	
Risk uncertainty	Predictability and clarity regarding risks and risk measures *(ISO standard length and share of standards under development)*
Potential for externalities	Risk severity & third-party exposure to harm when risks materialize *(National Risk Register: Impact and likelihood of risk)*
Public salience	Level of importance the public places on an industry's risks *(# search results on Google News across the last 5 years)*
Market size / concentration	Distribution of and total industry revenue across firms *(Herfindahl Index)*
Audit conditions	
Verification costs	Cost of establishing an auditee's conformity with rules *(Invasiveness of audit procedure)*
Information sensitivity	Potential harm from unauthorized use of information required for audit *(Governmental document sensitivity classifications)*
Skill specificity	Rarity and level of specialized expertise required for audit *(Level of market-based salary)*

Table 1: Definition and quantification proxy (in brackets) of demand-side factors. Each proxy value is categorized into high, medium and low for simplicity. Criticality refers to the first four factors.[5]

We posit that demand-side factors influence the trade-off between audit quality and audit efficiency. High levels

[4] Our quantification of capacity requirements of public bodies in section 7 is only a first, simplistic estimate. We acknowledge that this approach is imperfect as, e.g., the size of the AI industry does not necessarily correlate with the demand for AI audits, and depends on the jurisdiction.

[5] The demand-side factors are in line with Menard's three hybrid governance factors (2004), adapted for auditing. 1) Uncertainty is captured by risk uncertainty relating to the validity and reliability of information about risks. 2) Transaction costs are described on the extensive margin as the reasons for auditing transactions ("risk externalities"), on the intensive margin as the difficulty of the auditing transaction ("verification costs" and "information sensitivity"). 3) Asset specificity describes how the competence of auditors generalizes ("skill specificity"). However, economic hybrid governance theory is limited in focusing on economic hybridity and agents. This allows for this paper's actor-focused approach, but falls short of analyzing auditing from a within-organizational perspective (Bol et al. 2019) and power-distribution perspective (Levi-Faur 2011). To bridge the latter limitation, we build on regulation theorists (Behr 1985 and Stigler 1971) to establish factors which pertain to the existing power distribution, from a societal and an economic perspective. Following auditing scholars (Ramanna 2015), we separate power distribution into "public salience" and "market concentration and size".

of risk uncertainty, potential externalities, verification costs, and information sensitivity necessitate prioritizing audit quality, which is achieved through auditors' independence, competence, and access. Conversely, in large, less concentrated markets, audit efficiency becomes paramount, achieved through auditors' existing and adaptable capacity and relevant skills. These requirements for auditor characteristics subsequently dictate the allocation of audit responsibilities and the resources public bodies may need to develop.

4.2 Supply-Side Factors: Auditor Characteristics, Archetypes and Auditing Responsibilities

The factors outlined above define the demand for and challenges of auditing within an industry context. An appropriate auditing regime fulfills this demand by incentivizing independence and sufficient capacity (resources, competence and access) of auditors.

Supply-side factors (Audit characteristics)	
Independence	Absence of conflicts of interest (e.g. due to selection/payment by auditee), in public interest
Resources	Auditor's human, financial, computational and other resources; and flexibility of using resources
Competence	Auditor's skills and experiences in the kinds of audits demanded
Access	Extent of the auditor's access to evidence required for the audit (e.g., to data, tech, offices, staff)

Table 2: Definition of supply-side factors

We cluster different sets of auditor characteristics into four idealized auditor archetypes. In practice, a classification of public-appointed auditors as "highly independent" should be interpreted as a potential degree of independence, while currently many publicly-appointed auditors have conflicts of interest, e.g. due to simultaneous consulting work. Appendix C details our assumptions on auditor characteristics in depth.

Auditor type	**Auditor characteristics**			
	Independ.	**Resources**	**Competence**	**Access**
Public bodies	Public scrutiny	Inflexible	Built if salient	Clearances, mandates
Publicly-appointed	Quality for re-selection	Inflexible tendering	Specialized experts	Depends on security clearance
Auditee-selected	Lenient for re-selection	Flexible ecosystem		
Internal	Private interests	Directly available	Product-specific	Internal access

Level of auditor characteristics: *High* *Medium* *Low*

Table 3: Auditor archetypes[6] and potential, idealized characteristics suggested by the auditing literature. Auditors can build and change their characteristics (Details in Appendix C).

Auditing Responsibilities Along the Audit Lifecycle

We suggest that the lifecycle of all audit processes involves the following three stages (Raji et al. 2020, Ojewale et al. 2024)[7][8]:

1. **Developing** auditing methods and rules
2. **Collecting** evidence ('auditable artifacts') for the audit in accordance with the auditing method
3. **Judging** the evidence, producing an audit report.

The combination of audit scope (for advanced AI models: governance, security and model - see below) and audit lifecycle defines the auditing responsibility space. Different auditor archetypes can fulfill each responsibility. In the following, we observe who fulfills different responsibilities across case studies.

5 Auditing Regime Case Study Findings

5.1 Comparative Case Study Findings

The framework above is applied to each case, as illustrated below for one example case. In addition, each case is qualitatively examined along its historical emergence, responsibility setup and audit effectiveness. There are many factors shaping an auditing regime, like the degree of

[6] Given our focus on regulatory-demanded, statutory audits, we do not specifically list civil society or academic auditors - though the category of "publicly-appointed auditors" could be expanded to include them, while results remain similar.

[7] The demands of each stage depend on the type of audit undertaken and its purpose. Consider, for example, an AI model audit that utilizes benchmarking. Firstly, the auditor must select or develop the framework of benchmarks, resulting in a dedicated software package. This undertaking is technical and conceptual, requiring a match between the purpose of the audit and the metrics adopted. Multiple kinds of subject matter expertise may be required, e.g., relating to the model, auditing method and domain of interest for the audit (such as CBRN risks). Secondly, the auditor runs the AI model through the selected benchmarks to gather performance data. This may require substantial engineering effort, from preparing and formatting benchmark datasets through to ensuring benchmarking tools interface with the AI model. Thirdly, the auditor judges the performance data, decides on the need for additional tests and corrective measures, and produces the audit report.

[8] We exclude post-audit actions like transparency and enforcement considerations from this analysis for reasons of brevity. An audit of the auditor follows similar steps.

information access or continuity of audits (See Appendix A.1 for details on each case).

Case example: Cybersecurity audits in nuclear energy

Demand-side factors	Risk uncertainty	Medium *(<50% of ISO standards under development but >2000 pages)*
	Potential for external.	High *("catastrophic" classification)*
	Verification costs	Medium *(Inspections and simulations)*
	Information sensitivity	High *(Classification "restricted")*
	Market concentration	High *(Herfindahl Index of 1500)*
	Skill specificity	Medium *($122k salary for a nuclear cybersecurity analyst)*
	Public salience	High *(43M news search results)*
Supply-side	High criticality, thus independence important. High market concentration, thus inflexible capacity okay. High salience allows for capacity build-up in public bodies.	
Auditor	Who judges audit	Public bodies
	Who collects evidence	Public bodies & Internal
	Who develops audit	All
	Who audits the auditor	Public bodies

Table 4. Case example: Cybersecurity in nuclear energy.

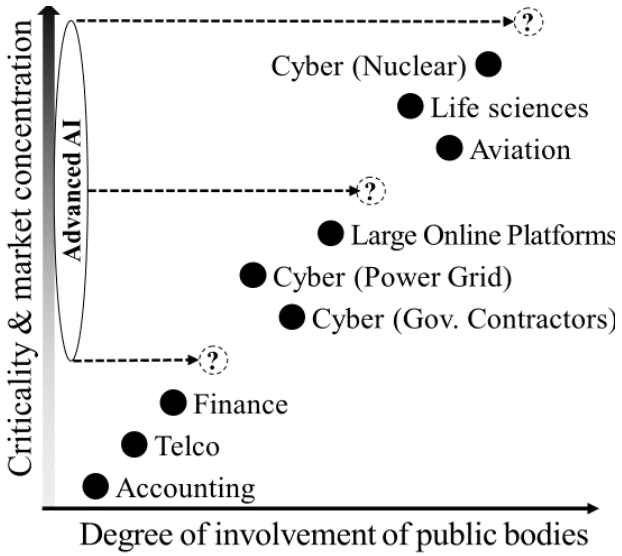

Figure 1: High criticality (risk externality, risk uncertainty, verification costs and info sensitivity) and market concentration of an industry is associated with relatively high involvement of public bodies in auditing (developing, collecting evidence, judging evidence and judging auditors). Both axes are quantified averages of the factors in brackets, for a typical product or security audit for each industry, as of 2024. Here they are displayed as ranks along the axes, thus distances between points are not meaningful. Details in Appendix A. As of 2024, advanced AI auditing by public bodies (-appointed) is limited (Hobbhahn and Scheurer 2024). Criticality of advanced AI is unclear.

The case studies illustrate that auditing regimes strike different compromises between independence and efficiency. Variation in regime design relates to demand-side characteristics in each context, such as risk uncertainty, the costs of verifying the safety of the audited technology, and the sensitivity of information uncovered during the auditing process. These factors positively correlate with the public body assuming greater control over the auditing process, prioritizing independence, safety and public trust over efficiency.

For nuclear energy cybersecurity and aviation safety, we empirically find a "critical" risk profile, and a high involvement of public bodies. However, it is not always effective for the public body to be highly involved in auditing. Intuitively, the more auditing that is demanded (because, for example, the market is larger and there are more audited firms), the more challenging it becomes for the public body to conduct each and every audit. For example, the diversity and quantity of radio frequency devices constrain the ability of the public body to conduct auditing in every instance. Similarly, the regime for public firms' annual reports requires more efficient auditing. Potential harm by radio frequency devices or accounting is relatively low, audit information less sensitive and verification possible without extensive trials. Thus, private parties are responsible for most auditing steps.

The following figure illustrates a potential explanation for different auditing responsibilities. Industry conditions and audit conditions (demand-side factors) demand different auditor characteristics (supply-side factors), essentially determining whether independence or efficient capacity are more important, which dictates who audits (auditing responsibility).

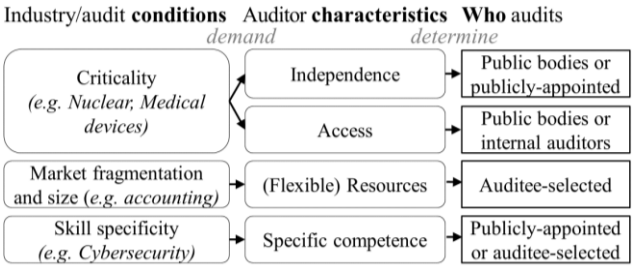

Figure 2. Connection between demand-side factors, supply-side factors and auditing responsibility. Note that auditor capacity can be influenced (see section 7). For each case, a combination of criticality, market concentration and skill specificity influences who audits, while criticality seems most prominent.

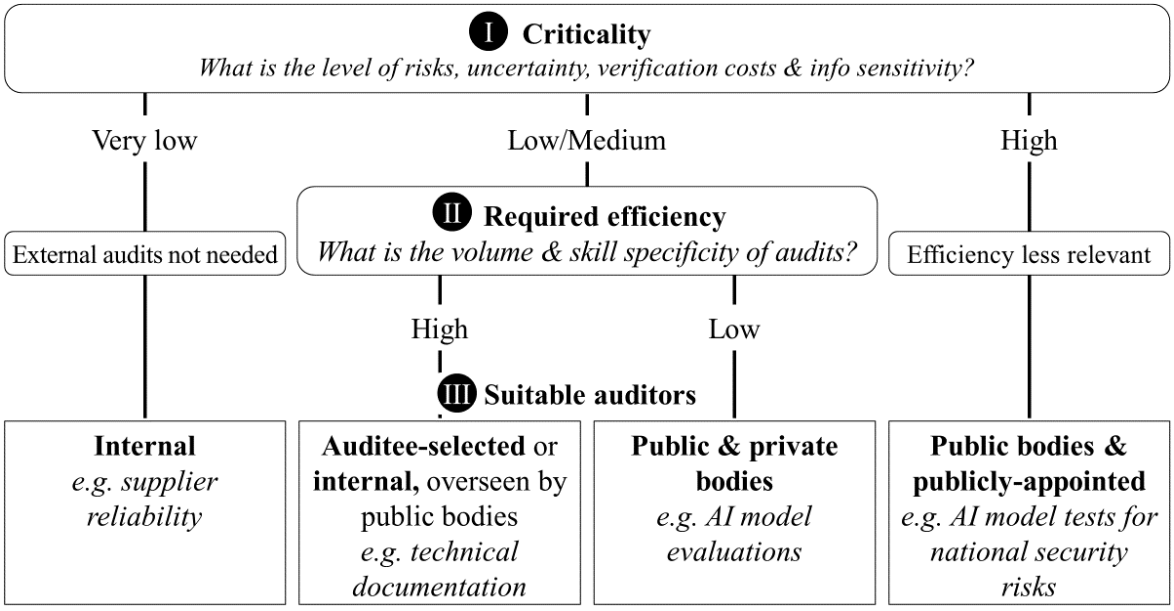

Figure 3: 3-step decision logic for running advanced AI auditing. Suitable auditors are indicative for collecting and judging evidence. The suitability is based on case study evidence on criticality and efficiency, and qualitatively explainable with auditor characteristics of independence and resources. Most likely, all auditor types might be involved in developing audits. AI Labs might support in all cases with collecting evidence. The volume of audits depends on future developments of market concentration and market size.

5.2 Three-Step Logic for Auditing Regime Design

Drawing on the quantification and analysis of cases above, we develop a three-step logic, intended to guide policymakers' auditing regime design choices (Figure 3).

- **Step 1 - Criticality**. Is the audit critical, necessitating an independent audit from a public body or publicly-appointed auditor? Criticality depends primarily on the risk level, risk uncertainty, verification costs and information sensitivity associated with the particular audit. It is only non-critical to involve auditee-selected auditors if the associated risks are well understood and the testing procedure is standardized.
- **Step 2 - Efficiency**. Who has or can efficiently build the required resources, competence and access? In this regard, we consider the volume of audits and the required skill specificity. If the volume of audits is high, and auditors do not require access to sensitive information, private parties may be tasked with auditing.
- **Step 3 - Suitable auditors**. Steps 1 and 2 determine which auditor characteristics are most demanded, and auditors with fitting characteristics are thus suitable.

This three-step logic is an idealized deduction from the case studies, reducing them to factors that previous literature and hybrid governance theory reasonably expects to influence audit effectiveness, as per our framework. However, the emergence of regimes is shaped by many other historical factors too, including political dynamics or concentration of skills in certain government departments (Ayres and Braithwaite 1992), as reviewed for each case in detail in Appendix A.1.

6 The Role of Public Bodies in an Advanced AI Auditing Regime

Public bodies can be involved in 6.1) different types of AI audits along different stages of the auditing lifecycle. The demand-side factors of each type determine 6.2) the optimal role of the public body in line with the three-step logic.

6.1 Advanced AI Audit Scope

There are many scopes or types of advanced AI audits. We focus on audits relevant to the development and provision of the AI model, and thus exclude product audits. We distinguish between those that focus on the governance practices of the firm that develops and provides advanced AI models, the security systems in place to prevent unauthorized access to the AI Lab's software and data, and

the capability, alignment and sociotechnical impacts of an AI model (Moekander et al. 2023, EU AI Act).

Governance audits ensure the firm meets structural and procedural prescriptions (Moekander et al. 2023, Crawford 2022). Governance audits are predominantly qualitative and examine documentation concerning the auditee's:

- Risk management system: risk identification, assessment, thresholds and mitigations, with emergency protocols in case of major incidents (Barrett et al. 2023)
- Quality management system: roles and responsibilities, points of contacts, system architecture, data governance
- Data audits (Birhane et al. 2024)
- Ecosystem audits: environmental reporting, labor, supply chain (Birhane et al. 2024)

Security audits evaluate the robustness of systems that prevent unauthorized access to and use of an AI Lab's technologies and data. They encompass assessments of cybersecurity systems, physical security systems, and information security systems (Nevo et al. 2023, Huang et al. 2024, Alaghbari et al. 2022).

Model audits evaluate AI models to explain their behaviors, assess their capabilities, and test their capacity for harm in user interactions and sociotechnical impacts (Weidinger et al. 2023, Casper et al. 2024). Black-box evaluation techniques assess an AI model's performance from an external (e.g., user) perspective, limiting analysis to the model's inputs and outputs without accessing its internal workings (Casper et al. 2024). By contrast, white-box techniques involve analyzing the internal functioning of the model (Casper et al. 2024). Intermediate approaches are referred to as 'gray-box'.

The required comprehensiveness of an audit may scale with an AI model's capabilities. For example, highly capable models, such as those trained with substantial computational resources, may require more rigorous audits. Common tiers of model audits include but are not limited to (OpenAI 2023, Anthropic 2024):

1. *Single-shot or few-shot benchmarking*. Evaluating the model's performance on specific tasks such as answering a set of multiple choice questions. There are different suites of benchmarks including the 'Measuring Multitask Language Understanding' (MMLU) measures, which test model accuracy on '57 tasks ranging from mathematics to history to law' (Anthropic, 2024, Liang et al. 2023)
2. *Black-box adversarial testing*. Technique aimed at intentionally exploiting a model to produce not intended outputs, such as an offensive image or instructions for cyberattacks. This may leverage domain-specific expertise, such as knowledge of chemical, biological, radiological and nuclear ("CBRN") threats (Anthropic 2023).
3. *Gray- or white-box, or scaffolding-enhanced adversarial tests*. Elicitation of capabilities and propensities of model behavior with extensive tooling on-top of the model or analysis of the internals of the model (Anthropic 2023).
4. *Systemic impact evaluations*, including human interaction evaluations, systemic safety monitoring, sociotechnical user studies, uplift studies and yet-to-be-developed audits of specific societal areas (Weidinger et al. 2023, Stein and Dunlop 2024).

This typology is not exhaustive. Other kinds of audits relevant to advanced AI are emerging such as code inspections (Cohen at al. 2024) and audits of computational resources (Sastry et al. 2024).

6.2 The Public Body's Optimal Role in an Advanced AI Auditing Regime

Below we apply the logic developed in Section 5 to the advanced AI context (Detailed sources: Appendix A.2.).

Demand-Side Analysis: Industry and Audit Factors

Industry conditions

Risk Uncertainty. Advanced AI is a complex and evolving general-purpose technology with implications for users and external systems that are expanding and difficult to reliably estimate, i.e. highly uncertain (DSIT 2024). There is a record number of 12 related standardization requests under discussions in JTC 21.

Potential for Externalities. Advanced AI already proliferates rapidly, with hundreds of millions of users worldwide (Stein and Dunlop 2024). The generality leads to an indefinite number of potential downstream use cases. The degree of risk externalities is debated and uncertain. In some scenarios, advanced AI poses catastrophic risks, in others, rather low externalities.

Public Salience. Currently, public salience of advanced AI risk is high (as measured by Google News results, see Appendix A.2), which allows for the build-up of public oversight capacity.

Market size and concentration. As a technology with high returns to scale, advanced AI model providers are highly concentrated. The 2024 generative AI industry size is $25 billion in the US (Statista 2024b). The industry is growing, but the audit volume remains highly uncertain.

Audit conditions

Verification Costs. Verifying the risks, safety and compliance of advanced AI systems can be complex and potentially costly, depending on the audit scope (see Table 5, and Brundage et al. (2020), Casper et al. (2024)). Current methods for adversarial tests, systemic impact analysis and security audits are unstandardized and require significant expertise, time, and resources, making thorough verification challenging. Other methods, like benchmarking, are less time intensive and more standardized.

Information Sensitivity. Adversarial model audits that identify flaws and vulnerabilities in highly capable AI models are sensitive to the extent they reveal pathways to misusing advanced AI for harmful purposes, like

cyberattacks or CBRN threats. There are also concerns that sensitive model test results could enter the training datasets of advanced AI. Due to the national security relevance of advanced AI, audits of security and AI models are sensitive. On the other hand, API-based black-box model evaluations need less sensitive information.

Skill Specificity. As foreshadowed in subsection 6.1, we suggest that particular AI model audits, as opposed to governance and security audits, require significant and specialized expertise. Domain-specific expertise is required to develop threat models and red-team advanced AI. Research engineers and computational social scientists are required to understand models and their impacts.

Audit scope	**Demand-side factors**		
	Industry risk profile (→*Resources*)	Skill specificity (→*Competence*)	Verification costs & info sensitivity (→*Access*)
Governance		E.g., Auditors in Compliance	E.g. Partly manual documentation
Security		E.g., Security professionals	E.g. Inspections, partly manual
Model...			
..Benchmarks		E.g., ML engineers	E.g. Black-box, automated
..Adversarial tests		E.g., Domain & ML experts	E.g. Grey-/White-box, manual
...Systemic impact		E.g., Social scientists	E.g. Black-box / Usage, manual

Level of demand-side factors: *High* *Medium* *Low*

Table 5: Assumed status quo of demand-side factors by audit scope, for advanced AI. Industry risk profile includes risk uncertainty, potential for externalities and market concentration. Access from Casper et al. (2024); competence based on practitioner input (see Appendix A).

Supply-Side Analysis: The Role of AI Safety Institutes and Other Public Bodies in Advanced AI Auditing

An advanced AI auditing regime should be designed to incentivize an optimal balance between the auditor's independence, resources, competence and access to auditing evidence. Failing this, we expect auditing quality and its usefulness as a tool for monitoring regulatory compliance and the benefits and safety of AI systems to deteriorate. The demand-side analysis of the industry risk profile suggests that the unpredictable but potentially critical and far-reaching impacts of advanced AI justify the prioritization of independence and, consequently, public body involvement. However, this finding is complicated by intersecting efficiency challenges of using existing and building new competence, access and capacity in a nascent and unstandardized ecosystem for AI model audits. Such audits require niche expertise and innovation in auditing practices. Therefore, as illustrated by Table 5, we suggest that different aims and types of AI audits invite different auditing regime considerations.

Implication 1: Public Body Involvement in Gray- and Black-Box Model Evaluations for Critical Risks

If policymakers agree with the demand-side analysis, we suggest that the public body should be directly involved in certain kinds of advanced AI model audits that: (a) pertain to critical risks such as those affecting national security; (b) demand white- or gray-box access to AI models (such as certain kinds of adversarial tests and evaluations); and (c) involve access to sensitive information. This model loosely resembles auditing regimes in aviation and nuclear energy. In line with the three-step logic, suitable auditors for such high criticality tasks are public bodies & publicly-appointed externals. Given concentration in the advanced AI market and the prohibitive costs of training state-of-the-art models, the volume of audits might allow for such high involvement of less efficient public bodies. However, as discussed in subsections 7.1 and 7.2 below, the challenge for policymakers is to ensure the public body possesses adequate expertise and knowledge of the advanced AI system to conduct intensive, complex and potentially bespoke model evaluations (Casper et al. 2024, Anthropic 2023). This could manifest as an integrated team comprising government officials, publicly-appointed experts and senior representatives from the AI Lab itself.

Implication 2: Public Oversight of an Auditing Market for Governance, Security and Select Model Audits

Governance and security audits of AI Labs are more standardized, tap into auditing practices that are relatively mature in other industry contexts, and, apart from certain kinds of security audits, do not entail access to information that would harm the public if disclosed (Schuett 2023, Bos 2018). We suggest, therefore, that such audits could be provided by a market of private auditors supplemented by public body oversight. The public body's role should be to facilitate high quality auditing through policies that augment auditor independence and competence. These should include schemes to accredit auditor expertise and regulations that impose quality standards on auditors with consequences for failure.

These considerations may also extend to certain kinds of black-box model audits such as benchmark evaluations that assess AI model performance on standardized tasks. Such evaluations do not typically involve highly sensitive information and could benefit from the competitive dynamics of a private auditing market, generating innovation and expertise in AI auditing practices.

7 Public Body Capacity Estimates

As a consequence of the criticality of some advanced AI audits and the necessity for government involvement analyzed in the previous chapter, public bodies like AI Safety Institutes must build regulatory capacity, technical competence and ensure information access to both conduct

certain kinds of audits and oversee others. In this section, we estimate the resources, competence and access requirements of the public body, with reference to case study evidence. Shortfalls in public bodies' capacity, competence and access limited effective AI auditing in the past (Lawrence et al. 2023, Groves et al. 2024, Politico 2024). Our figures are estimates only and assume the public body is operating in an advanced economy with a remit covering the current size of the advanced AI industry in the US.

When audits are critical: **More FTE at public bodies**

Criticality & market concentration ↑

	# FTE	# FTE (scaled)	Technical FTE	Info access
Cyber (Nuclear)	40	1250 – 2000	40-60%	On demand
Medical devices	1200	100-200	70-90%	On demand
Aviation	1800	600–900	70-90%	On demand
Advanced AI		?	?	?
Cyber (Power grid)	<30	250-400	70-90%	If suspicion
Finance (Securities)	30	<30	0-20%	Limited
Telco (Devices)	<30	<30	20-40%	Limited
Accounting	<30	<30	70-90%	If suspicion

Figure 5: Public bodies' resources across case studies in the US, sorted by criticality and market concentration. FTEs (Full-time equivalents) are scaled to the current advanced AI industry size of $25 billion in the US (Statista 2024b). Share of technical FTE are roles framed as "specialists". "Supportive" roles are non-technical staff. Info access on demand for random inspections (See details in Appendices A.3, B.4 and B.5)

7.1 Resources

Case Study Evidence

The cases suggest that in auditing regimes where the public body is directly involved in auditing, the public body employs more staff relative to when the public body is an overseer of private auditors. As analyzed previously, the public body is more involved in auditing, when criticality and market concentration are high. Thus, higher criticality and market concentration demands more staff at public bodies, as shown in Figure 5.

Implication 3: 100s of FTE for Advanced AI Auditing

For effective advanced AI auditing, public bodies' auditing-related FTEs, share of technical staff and access, will need to be roughly on par with public bodies active in other industries with similar criticality and market concentration. If criticality and market concentration of advanced AI remains high and thus demands high public involvement in auditing, then the public body will need 100s of auditing FTEs in jurisdictions like the US.

Advanced AI models or security audits that entangle sensitive information require precautionary measures to ensure evaluation results or test sets are not leaked publicly or introduced into the AI model's training data. We assume that particular model or security audits will be resource-intensive with up to a dozen auditors being required to collect and elicit evidence in respect of a single threat (see Appendix A.3 and B.4 for detailed estimations for each audit method).

A surge in new AI Labs, models and risks will require the public body to increase its auditing capacity. To adapt to changes in demand, public bodies may need to develop organizational slack (Bourgeois 1981) or flexibility by, for example, maintaining and drawing on a pool of accredited AI auditing experts from academia or private sectors. Framework agreements could assist in accelerating their appointment.

7.2 Competence

Case Study Evidence

What kind of staff are needed? We find that in auditing regimes where the public body is directly involved in specialist auditing methods related to a complex product or technology rather than corporate governance, a higher proportion of the public body's staff are technical specialists. For example, auditing teams in the US Food and Drug Administration ("FDA") are composed of >70% technical specialists, which reflects that the FDA is directly involved in assessing complex products such as medical devices. In regimes where the public body oversees a private auditing market, the public body is able to develop more generalist competence to, for example, focus on assessing auditors and processes rather than the safety and benefits of a technology itself.

Implication 4: Extensive, Diverse Technical Expertise at Public Bodies to Verify Claims of AI Labs

Required staff skill profiles vary depending on the specific audit. Public bodies require a mixed technical and non-technical team dedicated to developing, conducting or judging audits. This team should involve a mix of computer engineers, compliance specialists, and domain-specific experts from fields like cybersecurity (See Table 5). The share of technical profiles will depend on the public body's degree of involvement in auditing. As discussed, it seems likely that the public body will be very involved, at least in the medium-term, given high levels of risk uncertainty and a lack of standardization.

7.3 Access and Learning

Case Study Evidence

In addition to capacity and competence, auditors require access to the information necessary for auditing. When risk uncertainty and verification costs are high, public bodies and appointed auditors need extensive access to information held by auditees and auditors (Costanza-Chock et al. 2023).

Not only does sufficient access to information underpin auditing quality, it may also facilitate the development of auditing competence and standardization (Schelker 2010). However, private players who learn the most through internal access may not always have the incentive to share their learnings - as seen in the case of oil companies research on climate change or financial auditors' withholding of information as part of the Enron scandal (Petrick & Scherer 2003). Therefore, public bodies should ensure their learning through mandating access to: (A) auditees' information; and (B) auditors' information.

For technical, profit-aligned developments of audits, firms share information, speed up standardization and, in turn, increase innovation - like for telecommunications and cybersecurity in bulk power systems (Blind 2013, Blind 2006). In healthcare, cyber for government contractors, aviation, cyber for nuclear and life sciences, public bodies learned through continued information access, enabling more standardized guidelines and, over time, auditing by private auditors instead of directly by public bodies.

Implication 5: Structured Access to Auditee and Auditor Information

Verifying claims and conformance with rules of AI Labs requires structured access to facilities, security systems, and the AI model. Lacking access is a noted challenge for AI auditors (Costanza-Chock et al. 2023, Casper et al. 2024). To effectively collect and judge evidence, e.g. by conducting in-depth evaluations and adversarial tests, gray- and white-box access might be required (see Figure 6). For API-based benchmarks or developing audits access to proxies and analogous samples (i.e., sufficiently similar but not identical datasets) may suffice. Systemic impact and human interaction evaluations might require access to anonymized usage or human trial data (Weidinger et al. 2023).

Given the current concentration of expertise and the need to swiftly develop (harmonized) standards in advanced AI, public bodies and trusted researchers need access to private sector expertise and information.

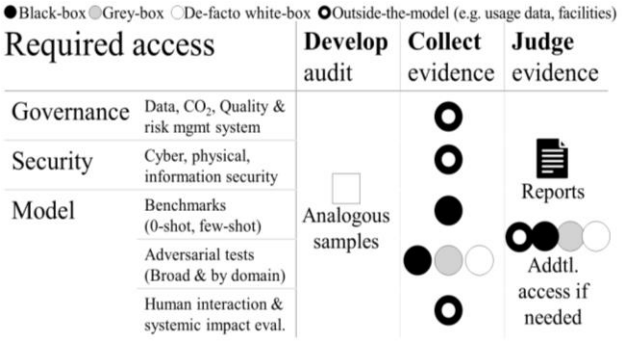

Figure 6: Access for auditing advanced AI. Terminology based on Casper et al. (2024)

When risks are more certain and audits standardized, auditee information can be restricted to cases of suspicion. AI audits that identify flaws and vulnerabilities in highly capable models may reveal pathways to misusing advanced AI for harmful purposes. Consequently, policymakers must mandate the optimal level of information access for AI auditors, instigating safeguards such as the requirement to obtain security clearances for gray- and white-box audits of highly capable AI models.

8 Conclusion

Drawing on our analysis of auditing regimes across high-risk industries, we derived five implications for designing advanced AI auditing regimes. Implications 1 and 2 revealed that when advanced AI risks, risk uncertainty, verification costs and information sensitivity are at levels comparable to the nuclear energy or aviation sectors, public bodies and publicly-appointed specialists need to audit AI Labs directly. Implications 3, 4, and 5 described the required resources, competence and access for AI Safety Institutes and public bodies to fulfill their auditing role. In case of high criticality, 100s of sociotechnical FTE and structured access to auditee and auditor information are needed.

Future research could explore a wider range of auditing regimes and country contexts, using deductive methodologies to test findings. Such research might, for example, consider:

- Quantitatively investigating causal links between auditing regime design choices and regime effectiveness.
- Qualitatively describing nuanced dynamics within and between AI auditor organizations (both public and private), exploring, for example, power dynamics, regulatory capture, and cultural differences.
- Understand the historical political and institutional reasons how different regimes and public bodies developed best practices and standardized audits.
- Define how auditing intersects with other AI governance mechanisms as part of a comprehensive regime.

Acknowledgments

We are grateful for helpful discussions and feedback from Aaron Maniam, Alan Chan, Alexis Carlier, Clíodhna Ní Ghuidhir, Elliot Jones, Friederike Grosse-Holz, Gaurav Sett, Herbie Bradley, Lewis Ho, Lisa Soder, Lujain Ibrahim, Patrick Levermore, Peter Wills, Roxana Radu and participants of multiple AI Governance workshops in Oxford.

Appendices

Available on: https://arxiv.org/abs/2407.20847

References

Alaghbari, K. A.; Saad, M. H. M.; Hussain, A.; and Alam,

M. R. 2022. Complex event processing for physical and cyber security in datacentres - recent progress, challenges and recommendations. Journal of Cloud Computing, 11(1): 65.

Allied Market Research, A. M. R. 2023. Cyber Security in Energy Sector Market Size, Forecast - 2032.

Anderljung, M.; Smith, E. T.; O'Brien, J.; Soder, L.; Bucknall, B.; Bluemke, E.; Schuett, J.; Trager, R.; Strahm, L.; and Chowdhury, R. 2023. Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework. ArXiv:2311.14711 [cs].

Anthropic. 2023. Challenges in evaluating AI systems \ Anthropic.

Anthropic. 2024. Model Card Claude 3.pdf.

Ayres, I.; and Braithwaite, J. 1992. Responsive Regulation: Transcending the Deregulation Debate. New York: Oxford University Press. Print.

Barrett, Anthony M.; Newman, Jessica; Nonnecke, Brandie; Hendrycks, Dan; Murphy, Evan R.; and Jackson, Krystal. 2023. Berkeley-GPAIS-Foundation-Model-Risk- Management-Standards-Profile-v1.0.pdf.

Behr, R. L.; and Iyengar, S. 1985. Television News, Real- World Cues, and Changes in the Public Agenda. Public Opinion Quarterly, 49(1): 38–57.

Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. ArXiv:2401.14462 [cs].

Blind, K. 2006. Explanatory factors for participation in formal standardisation processes: Empirical evidence at firm level. Economics of Innovation and New Tech- nology, 15(2): 157–170. Publisher: Routledge eprint: https://doi.org/10.1080/10438590500143970.

Blind, K. 2013. The Impact of Standardization and Standards on Innovation.

Bos, G. 2018. ISO 13485:2003/2016—Medical Devices—Quality Management Systems—Requirements for Regulatory Purposes. In Handbook of Medical Device Regulatory Affairs in Asia. Jenny Stanford Publishing, 2 edition. ISBN 978-0-429-50439-6. Num Pages: 22.

Bourgeois, L. J. 1981. On the Measurement of Organizational Slack. The Academy of Management Review 6(1): 29–39. doi.org/10.2307/257138. Accessed: 2024-07-25.

OECD 2014. Measuring Environmental Policy Stringency in OECD Countries: A Composite Index Approach. Technical report, OECD, Paris.

Broecke, S. 2016. Do skills matter for wage inequality? IZA World of Labor.

Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R.; Maharaj, T.; Koh, P. W.; Hooker, S.; Leung, J.; Trask, A.; Bluemke, E.; Lebensold, J.; O'Keefe, C.; Koren, M.; Ryffel, T.; Rubinovitz, J. B.; Besiroglu, T.; Carugati, F.; Clark, J.; Eckersley, P.; de Haas, S.; Johnson, M.; Laurie, B.; Ingerman, A.; Krawczuk, I.; Askell, A.; Cammarota, R.; Lohn, A.; Krueger, D.; Stix, C.; Henderson, P.; Graham, L.; Prunkl, C.; Martin, B.; Seger, E.; Zilberman, N.; hE´ igeartaigh, S. ; Kroeger, F.; Sastry, G.; Kagan, R.; Weller, A.; Tse, B.; Barnes, E.; Dafoe, A.; Scharre, P.; Herbert-Voss, A.; Rasser, M.; Sodhani, S.; Flynn, C.; Gilbert, T. K.; Dyer, L.; Khan, S.; Bengio, Y.; and Anderljung, M. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. ArXiv:2004.07213 [cs].

Casper, S.; Ezell, C.; Siegmann, C.; Kolt, N.; Curtis, T. L.; Bucknall, B.; Haupt, A.; Wei, K.; Scheurer, J.; Hobb- hahn, M.; Sharkey, L.; Krishna, S.; Von Hagen, M.; Al- berti, S.; Chan, A.; Sun, Q.; Gerovitch, M.; Bau, D.; Tegmark, M.; Krueger, D.; and Hadfield-Menell, D. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. ArXiv:2401.14446 [cs].

Coase, R. H. 1960. The Problem of Social Cost. The Journal of Law & Economics, 3: 1–44. Publisher: [University of Chicago Press, Booth School of Business, University of Chicago, University of Chicago Law School].

Code of Federal Regulations Title 14, U. S. 2024. 14 CFR Part 21 – Certification Procedures for Products and Articles.

Code of Federal Regulations Title 16, U. S. 2024. eCFR :: 16 CFR Part 314 – Standards for Safeguarding Customer Information.

Code of Federal Regulations Title 47, U. S. 2024. 47 CFR § 68.162 Requirements for Telecommunication Certification Bodies.

Coherent Market Insights, U. S. 2024. Defense Cyber Security Market - Price, Size, Share & Growth.

Cohen, M. K.; et al. 2024. Regulating Advanced Artificial Agents. Science 384: 36–38. doi.org/10.1126/science.adl0625.

Costanza-Chock, S.; Harvey, E.; Raji, I. D.; Czernuszenko, M.; and Buolamwini, J. 2023. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In 2022 ACM Conference on Fairness, Accountability, and Transparency, 1571–1583. ArXiv:2310.02521 [cs].

Crawford. 2022. Atlas of AI.

Cyber AB, U. S. 2024. The Cyber AB: Overview.

Davenport, C. 2023. SpaceX to the FAA: The industry needs you to move faster. Washington Post.

DeFond, M. L. 2010. How should the auditors be audited? Comparing the PCAOB Inspections with the AICPA Peer Reviews. Journal of Accounting and Economics, 49(1): 104–108.

DeFond, M.; and Zhang, J. 2014. A Review of Archival Auditing Research. Journal of Accounting and Economics 58(2–3): 275–326. doi.org/10.1016/j.jacceco.2014.09.002.

Dennis, K. 2008. The Rating Game: Explaining Rating Agency Failures in the Buildup to the Financial Crisis. University of Miami Law Review, 63(4): 1111–1150.

Department of Defense, U. S. 2024. About CMMC.

Department of Science, Innovation and Technology (DSIT), U.K Government. 2024. International Scientific Report on the Safety of Advanced AI.

Digital Services Act, E. 2022.

Duflo, E.; Greenstone, M.; Pande, R.; and Ryan, N. 2013. Truth-telling by Third-party Auditors and the Response of Polluting Firms: Experimental Evidence from India*. The Quarterly Journal of Economics, 128(4): 1499–1545. Publisher: Oxford Academic.

Efing, M.; and Hau, H. 2015. Structured debt ratings: Evidence on conflicts of interest. Journal of Financial Economics, 116(1): 46–60.

EpochAI, E. 2023. Machine Learning Trends.

Bol, A., Grabner, I., Vienna, W., & Haesebrouck, K. 2019. Literature review The effect of audit culture on audit quality. Found Audit Res.

EU. 2022. The Digital Services Act (DSA) - Regulation (EU) 2022/2065.

European Commission, E. 2018. DocsRoom European Commission.

European Commission, E. 2024. Do you want to help enforce the Digital Services Act? Apply now to be part of the DSA enforcement team! | Shaping Europe's digital future.

European Council, E. 2024. The EU's platform economy.

European Court of Auditors, E. 2015. EU supervision of credit rating agencies – well established but not yet fully effective. credit rating agencies.

European Union Aviation Safety Agency, E. 2024. Aircraft certification | EASA.

FDA 2023. CLIA Categorizations. FDA. Publisher: FDA.

Federal Aviation Administration, U. S. 2021a. Aviation Safety Workforce Plan 2021 | 2030.

Federal Aviation Administration, U. S. 2021b. A Brief History of the FAA | Federal Aviation Administration.

Federal Aviation Administration, U. S. 2021c. Legal Enforcement Actions | Federal Aviation Administration.

Federal Aviation Administration, U. S. 2022. Airworthiness Certification Overview | Federal Aviation Administration.

Federal Aviation Administration, U.S. 2022. Bilateral Agreements | Federal Aviation Administration.

Federal Communications Commission, U. S. 2015. Laboratory Division | Federal Communications Commission.

Federal Communications Commission (webpage). Equipment Authorization – RF Device | Federal Communications Commission.

Federal Communications Commission (webpage). Equipment Authorization Procedures | Federal Communications Commission.

Federal Communications Commission, U. S. 2023. Office of Engineering and Technology (OET) Organization Chart

| Federal Communications Commission.

Federal Communications Commission, U. S. 2024a. DOC-391605A1.pdf.

Federal Communications Commission, U. S. 2024b. Enforcement Primer | Federal Communications Commission.

Federal Communications Commission, U. S. 2024c. Equipment Authorization | Federal Communications Commission.

Federal Communications Commission, U. S. 2024d. Main webpage (William H. Donaldson).

Federal Energy Regulatory Commission, U. S. 2022. Career Opportunities | Federal Energy Regulatory Commission.

Federal Energy Regulatory Commission, U. S. 2023a. FERC FY 24 Congressional Justification | Federal Energy Regulatory Commission.

Federal Energy Regulatory Commission, U. S. 2023b. Re- lated Document Classes | Federal Energy Regulatory Commission.

Federal Energy Regulatory Commission, U. S. 2023c. Reliability Explainer | Federal Energy Regulatory Commission.

Financial Industry Regulatory Authority, U. S. 2010. 4140. Audit | FINRA.org.

Financial Industry Regulatory Authority, U. S. 2024. What We Do | FINRA.org.

Fiolleau, K.; Hoang, K.; Jamal, K.; and Sunder, S. 2013. How Do Regulatory Reforms to Enhance Auditor Independence Work in Practice? Contemporary Accounting Research, 30(3): 864–890. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1911-3846.12004.

Food and Drug Administration, U. S. 2018. Details of Full- Time Equivalents.

Food and Drug Administration, U. S. 2023. Jobs at the Center for Devices and Radiological Health (CDRH). Publisher: FDA.

Food and Drug Administration, U. S. 2024a. 510(k) Third Party Review Program. Publisher: FDA.

Food and Drug Administration, U. S. 2024b. CDRH Man- agement Directory by Organization. FDA. Publisher: FDA.

Food and Drug Administration, U. S. 2024c. FDA's Risk- Based Approach to Inspections. FDA. Publisher: FDA.

Food and Drug Administration, U. S. 2024d. A History of Medical Device Regulation & Oversight in the United States. FDA. Publisher: FDA.

Food and Drug Administration, U. S. 2024e. Overview of Device Regulation. Publisher: FDA.

Fortune Business Insights, U. S. 2024. U.S. Medical Devices Market Size, Share | Analysis Report, 2030.

Galland, J.-P. 2024. Standards, Certification, and Accreditation: Indispensable Tools for European Safety Regulations? In The Regulator–Regulatee Relationship in High-Hazard Industry Sectors, 71–78. Springer, Cham. ISBN 978-3-031- 49570-0. ISSN: 2191-5318.

Glassdoor. 2024. Company Salaries.

Government Accountability Office, U. S. 2021. Securities Regulation: SEC Could Take Further Actions to Help Achieve Its FINRA Oversight Goals | U.S. GAO.

Groves, L.; Metcalf, J.; Kennedy, A.; Vecchione, B.; and Strait, A. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. ArXiv:2402.08101 [cs].

Gunningham, N.; Grabosky, P.; and Sinclair, D. 1998. Smart Regulation: Designing Environmental Policy. Oxford: Oxford Academic. doi.org/10.1093/oso/9780198268574.001.0001.

Hadfield, G.; Clark, J. 2023. Regulatory Markets: The Future of AI Governance. ArXiv:/2304.04914 [cs]

Hansen, S. C.; Kumar, K. R.; and Sullivan, M. W. 2008. Auditor Capacity Stress and Audit Quality: Market-Based Evidence from Andersen's Indictment.

Hazlett, T.; and Pai, A. 2018. The Untold History of FCC Regulation.

Hill, M.; and Varone, F. 2021. The Public Policy Process. London: Routledge, 8 edition. ISBN 978-1-00-301020-3.

Hobbhahn, M.; and Scheurer, J. 2024. Apollo Research. We need a Science of Evals.

Huang, K.; Wang, Y.; Goertzel, B.; Li, Y.; Wright, S.; and Ponnapalli, J., eds. 2024. Generative AI Security: Theories and Practices. Future of Business and Finance. Cham: Springer Nature Switzerland. ISBN 978-3-031-54251-0 978-3-031-54252-7.

IEEE. 2008. IEEE Standard for Software Reviews and Audits. ISBN: 9780738157689.

ISO/IEC. 2015. ISO/IEC 17021-1:2015 - Conformity assessment - Requirements for bodies providing audit and certification of management systems - Part 1: Requirements.

ISO/IEC. 2022. ISO/IEC 27001:2022 - Information security, cybersecurity and privacy protection — Information security management systems — Requirements.

Jorgenson, D. W.; Landefeld, J. S.; and Schreyer, P. 2014. Measuring Economic Sustainability and Progress. Univer- sity of Chicago Press. ISBN 978-0-226-12133-8 978-0-226- 12147-5.

Kleinman, G.; Lin, B. B.; and Palmon, D. 2014. Audit Quality: A Cross-National Comparison of Audit Regulatory Regimes. Journal of Accounting, Auditing & Finance, 29(1): 61–87. Publisher: SAGE Publications Inc.

Kowaleski, Z. T.; Mayhew, B. W.; and Tegeler, A. C. 2018. The Impact of Consulting Services on Audit Quality: An Experimental Approach. Journal of Accounting Research, 56(2): 673–711. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-679X.12197.

Kurt, A. C. 2022. Audit Risk and the Implications of Employing Specialist Auditors: Evidence from Government Contractors.

Lamoreaux, P. T. 2016. Does PCAOB inspection access improve audit quality? An examination of foreign firms listed in the United States. Journal of Accounting and Economics, 61(2): 313–337.

Lawrence, C.; Cui, I.; and Ho, D. 2023. The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 606–652. Montr\'{e}al QC Canada: ACM. ISBN 9798400702310.

Lennon, H. 2021. Why The SEC's Stance On Bitcoin ETFs May Need To Change. Section: Crypto & Blockchain.

Levi-Faur, D. 2003. Comparative Research Designs in the Study of Regulation: How to Increase the Number of Cases Without Compromising the Strengths of Case- Oriented Analysis.

Levi-Faur, D. 2011. Regulation and Regulatory Governance. In Handbook on the Politics of Regulation, edited by D. Levi-Faur, Chapter 1. Cheltenham, UK: Edward Elgar Publishing.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Re´, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L.; Zheng, L.; Yuksekgonul, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2023. Holistic Evaluation of Language Models. ArXiv:2211.09110 [cs].

Marco, A. C.; Carley, M.; Jackson, S.; and Myers, A. 2017. The USPTO Historical Patent Data Files: Two Centuries of Innovation.

Maynard, M. 2014. The FCC TCB Program: A Government and Industry Cooperative.

MDC. 2022. Price List (Certification according to MDR).

Mellon, J. 2013. Internet Search Data and Issue Salience: The Properties of Google Trends as a Measure of Issue Salience. Journal of Elections, Public Opinion and Parties, 24(1): 45–72. Publisher: Routledge eprint: https://doi.org/10.1080/17457289.2013.846346.

METR. 2024. Portable Evaluation Tasks via the METR Task Standard.

Moore, D.; Tanlu, L.; and Bazerman, M. 2010. Conflict of Interest and the Intrusion of Bias. Judgment and Decision Making, 5: 37–53.

Mutchler, J. F. 2003. Chapter 7: Independence and Objectivity - A Framework for Research Opportunities in Internal Auditing.

Menard, C. 2004. The Economics of Hybrid Organizations. Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift fu¨r die gesamte Staatswissenschaft, 160(3): 345–376. Publisher: Mohr Siebeck GmbH & Co. KG.

Moekander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing large language models: a three-layered approach. AI and Ethics.

National Institute of Standards and Technology, U. S. 2023. Staff. NIST.

National Institute of Standards and Technology, U. S. 2024a. Accreditation Programs. NIST. Last Modified: 2023-05-17T10:59-04:00.

National Institute of Standards and Technology, U. S. 2024b. Government Contractor Requirements. NIST. Last Modified: 2024-02-07T10:09-05:00.

National Institute of Standards and Technology, U. S. 2024c. Telecommunications Certification Bodies (TCB) Application Information. NIST. Last Modified: 2021-06-02T18:27- 04:00.

National Risk Register, U. K. 2023. 2023 NATIONAL RISK REGISTER NRR.pdf.

National Science and Technology Council, U.S. Government. 2024. Critical and Emerging Technology List Update.

Neuman, J. 2008. FAA's 'culture of coziness' targeted in airline safety hearing. Section: Travel & Experiences.

Nevo, S.; Lahav, D.; Karpur, A.; Alstott, J.; and Matheny,

J. 2023. Securing Artificial Intelligence Model Weights: Interim Report. RAND Corporation.

Nextlabs.2016. SB-NERC-and-FERC-Cyber-Security.pdf.

Noll, R. G. 1989. Max D. Paglin, ed., A legislative history of the communications Act of 1934: (Oxford University Press, New York, 1989) pp. xxii+981. Information Economics and Policy, 4(2): 190–194.

North American Electric Reliability Corporation, U. S. 2023. Reliability Principles.pdf.

NQA. 2024a. AS Certifications - AS9100 / AS9110 / AS9120 / AS6081 | NQA.

NQA. 2024b. AS9100 Certification - Aerospace Management Standard | NQA.

Nuclear Regulatory Commission, U. S. 2021a. Information Security.

Nuclear Regulatory Commission, U. S. 2021b. Opportunities.

Nuclear Regulatory Commission, U. S. 2021c. § 73.54 Protection of digital computer and communication systems and networks.

Nuclear Regulatory Commission, U. S. 2024a. Cybersecurity | NRC.gov.

Nuclear Regulatory Commission, U. S. 2024b. ML24061A093.pdf.

Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. ArXiv:2402.17861 [cs].

OpenAI. 2023. gpt-4-system-card.pdf.

Pasztor. 2023. Opinion: The FAA's Safety System Is Starting to Show Cracks | Aviation Week Network.

Pawlson, L. G.; Torda, P.; Roski, J.; and O'Kane, M. E. 2005. The role of accreditation in an era of market-driven accountability. The American Journal of Managed Care, 11(5): 290–293.

PCAOB. 2024. Standards | PCAOB.

Petrick, J. A.; and Scherer, R. F. 2003. The Enron Scandal and the Neglect of Management Integrity Capacity. American Journal of Business, 18(1): 37–50. Publisher: MCB UP Ltd.

Pigou. 1920. The Economics of Welfare.

Politico. 2024. Rishi Sunak promised to make AI safe. Big Tech's not playing ball.

Power, M. 1999. The Audit Society: Rituals of Verification. Oxford University Press. ISBN 978-0-19-168518-7.

Quélin, B. V.; Cabral, S.; Lazzarini, S.; and Kivleniece, I. 2019. The Private Scope in Public–Private Collaborations: An Institutional and Capability-Based Perspective. Organization Science, 30(4): 831–846. Publisher: INFORMS.

Radu, R. 2021. Steering the governance of artificial intelligence: national strategies in perspective. Policy and Society, 40(2): 178–193.

Rajala, T.; and Kokko, P. 2021. Biased by design – the case of horizontal accountability in a hybrid organization. Accounting, Auditing & Accountability Journal, 35(3): 830– 862. Publisher: Emerald Publishing Limited.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: defining an endto-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. Barcelona Spain: ACM. ISBN 978-1-4503-6936-7.

Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. E. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. ArXiv:2206.04737 [cs].

Ramanna, K. 2015. Thin Political Markets: The Soft Underbelly of Capitalism. California Management Review, 57(2): 5–19. Publisher: SAGE Publications Inc.

ReedSmith. 2024. Hiring Non-U.S. Citizens – Don't Forget to Get Your Export Licenses | Perspectives | Reed Smith LLP.

Precedence Research. 2023. Aerospace Market Size To Reach USD 678.17 Billion By 2032.

Precedence Research 2024. RF Components Market Size To Hit USD 101.09 Bn By 2032.

Princeton UniversityUniversity, P. 2024. Sharing Technical Information or Software.

Ryan, J. 2012. The Negative Impact of Credit Rating Agencies and proposals for better regulation.

Sastry, G.; Heim, L.; Belfield, H.; Anderljung, M.; Brundage, M.; Hazell, J.; O'Keefe, C.; Hadfield, G. K.; Ngo, R.; Pilz, K.; Gor, G.; Bluemke, E.; Shoker, S.; Egan, J.; Trager, R. F.; Avin, S.; Weller, A.; Bengio, Y.; and Coyle, D. 2024. Computing Power and the Governance of Artificial Intelligence. ArXiv:2402.08797 [cs].

S&P Global RatingsRatings, S. G. 2022. guide to credit rating essentials digital.pdf.

Schelker, M. 2010. Auditor Expertise: Evidence from the Public Sector.

Schuett, J. 2023. Three lines of defense against risks from AI. AI & SOCIETY. ArXiv:2212.08364 [cs].

Securities and Exchange Commission, U. S. 2020a. EJR Ex 2 Methodologies.pdf.

Securities and Exchange Commission, U. S. 2020b. SEC.gov | The SEC's Office of Credit Ratings and NRSRO Regulation: Past, Present, and Future.

Securities and Exchange Commission, U. S. 2022. SEC.gov | Types of Appointment Authorities.

Securities and Exchange Commission, U. S. 2023. fy-2024-congressional-budget-justification final-3-10.pdf.

Short, J. L.; Toffel, M. W.; and Hugill, A. R. 2016. Monitoring global supply chains. Strategic Management Journal, 37(9): 1878–1897. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smj.2417.

Simnett, R.; Carson, E.; and Vanstraelen, A. 2016. International Archival Auditing and Assurance Research: Trends, Methodological Issues, and Opportunities. AUDITING: A Journal of Practice & Theory, 35(3): 1–32.

Simnett, R.; and Trotman, K. T. 2018. Twenty-Five-Year Overview of Experimental Auditing Research: Trends and Links to Audit Quality. Behavioral Research in Accounting, 30(2): 55–76.

Solomon, S. D. 2010. The Government's Elite and Regulatory Capture. Section: Business Day.

Statista. 2023a. Cybersecurity - United States | Statista Market Forecast.

Statista. 2023b. Topic: Accounting industry in the U.S.

Statista. 2024a. Forecast: Industry revenue of "securities brokerage" in the U.S. 2012-2024.

Statista. 2024b. Generative AI - North America | Statista Market Forecast.

Stein, M.; and Dunlop, C. 2023. Safe before sale.

Stein, M.; and Dunlop, C. 2024. Safe beyond sale: Post-deployment monitoring of AI.

Stigler, G. J. 1971. The Theory of Economic Regulation. The Bell Journal of Economics and Management Science 2(1): 3–21. doi.org/10.2307/3003160. Accessed: 2024-07-25.

Talent. 2024. Salary in USA - Average Salary.

Tanner, B. 2000. Independent assessment by third-party certification bodies. Food Control, 11(5): 415–417.

The White House, U. S. 2009. Executive Order 13526. Classified National Security Information.

The White House, U. S. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

Tortoise. 2023. The Global AI Index - Tortoise.

UK Government. 2023. Safety and security risks of generative artificial intelligence to 2025 (Annex B).

Van Loo, C. L. 2019. Regulatory Monitors: Policing Firms in the Compliance Era.

Watkins, L. . 2022. SEC Investigations.

Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; Gabriel, I.; Rieser, V.; and Isaac, W. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. ArXiv:2310.11986 [cs].

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; Biles, C.; Brown, S.; Kenton, Z.; Hawkins, W.; Stepleton, T.; Birhane, A.; Hendricks, L. A.; Rimell, L.; Isaac, W.; Haas, J.; Legassick, S.; Irving, G.; and Gabriel, I. 2022. Taxonomy of Risks posed by Language Models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, 214–229. New York, NY, USA: Association for Computing Machinery. ISBN 978-1- 4503-9352-2.

White, L. J. 2018. The Credit Rating Agencies and Their Role in the Financial System. Book Title: The Oxford Handbook of Institutions of International Economic Governance and Market Regulation Edition: 1 ISBN: 9780190900571 9780190900601 Publisher: Oxford University Press