How could frontier AI developers implement internal audit?

We explore how choices around scope, sourcing, frequency, and information access shape whether internal audit can credibly assure frontier AI risks without slowing development or exposing sensitive information.

By Francesca Gomez, Adam Buick, Leah Ferentinos, Haelee Kim and Elley Lee

Read full paper here.

As AI systems become more capable, the risks they pose grow more serious. The most advanced AI models can now assist with tasks that could be dangerous in the hands of certain actors - for example, the development of bioweapons or the automation of cyber attacks. As the capability for advanced AI to lead to harm increases, it becomes increasingly important that the organizations developing these models are managing their risks effectively.

Discussions of AI safety often focus on whether AI companies are assessing dangerous model capabilities and putting mitigations in place at the model level. But as AI systems cross capability levels for chemical, biological, radiological, and nuclear (CBRN) weapons development, safety increasingly depends on controls that sit outside the models themselves, such as how securely model weights are handled, how effectively incidents are detected and responded to, and whether governance structures give decision-makers an accurate view of what is actually happening across models and operational processes.

This raises a crucial question: how can boards and other stakeholders gain assurance that an AI company's governance, risk management, and control processes are genuinely effective in practice? And can this be done in a way which accommodates the speed of frontier AI development and sensitivities about sharing information?

In a new paper, we examine a well-established assurance solution from other high-stakes industries: **internal audit**. For decades, institutions such as banks have used internal audit to provide independent, evidence-based assurance that risks are properly managed. Building on work by Schuett (2024), who presents the case for frontier AI developers implementing an internal audit function to provide assurance, we explore the practical options for this, examining the trade-offs around what to audit, who should conduct it, how frequently, and what information auditors require.

Motivation

A further motivation for examining the practical feasibility of assurance options for frontier AI developers comes from the EU General-Purpose AI Code of Practice. Commitment 8.1 in the Safety and Security chapter includes a requirement for signatories to allocate responsibility for providing assurance about the adequacy of systemic risk assessment and mitigation processes to the board or equivalent governance body.

> **Systemic risk assurance:** The responsibility for providing assurance about the adequacy of the Signatory's systemic risk assessment and mitigation processes and measures to the management body in its supervisory function or another suitable independent body (such as a council or board) has been assigned to a relevant party (e.g. a Chief Audit Executive, a Head of Internal Audit, or a relevant sub-committee). This individual is supported by an internal audit function, or equivalent, and external assurance as appropriate. The Signatories' internal assurance activities are appropriate. For Signatories that are SMEs or SMCs, the management body in its supervisory function periodically assesses the Signatory's systemic risk assessment and mitigation processes and measures (e.g. by approving the Signatory's Framework assessment).

Measure 8.4, part (4): https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

In practice, this means designating someone to independently review how well the organization is managing systemic risks, examining evidence of how this is being done, and providing evidence-backed statements to the board that can be relied upon for governance decisions.

The core design choices for an internal audit function

Implementing an internal audit function requires strategic decisions across four key dimensions. The table below summarizes the options available to frontier AI developers:

Question	Options
What risks should internal audit provide assurance about?	• Model-level risks (e.g., dangerous capabilities, evaluation validity, post-training safeguards, refusal systems). • System-level risks (e.g., access controls, logging, monitoring, incident response, secure compute, weight protection). • Governance-level risks (e.g., board oversight, safety committees, release gates, escalation, override controls).
How should the internal audit function be sourced?	• Fully in-house internal audit team. • Co-sourced model (internal Chief Audit Executive + external specialists). • Fully outsourced to an external provider.
How frequently should internal audits be conducted?	• Annual cycles. • Semi-annual or quarterly cycles. • Ad-hoc reviews / rapid assurance. • Continuous auditing.
What information sources should internal auditors access?	• Structural information (e.g., governance frameworks, safety policies, organisation charts). • Procedural information (e.g., internal reports, system documentation, control designs). • Operational information (e.g., staff interviews, incident records, internal testing outputs). • Technical information (e.g., model evaluations, logs, monitoring systems, codebases, infrastructure).

Choices from the above options will be largely shaped by two key considerations: risk-based priorities determining which areas are most critical to have assurance over and practical constraints.

Risk-based priorities are typically determined through a process called risk-based auditing which results in an internal audit plan, a schedule of functions, processes, systems, and governance arrangements to be reviewed over the coming months. This plan is proposed by the Chief Audit Executive (CAE) and approved by the board based on which areas are most critical to have assurance over. It determines the information access required for reasonable assurance and specifies audit frequency to maintain ongoing coverage. Practical constraints which could shape the final agreed plan include the need to avoid overburdening teams with excessive auditing and restrictions on sharing sensitive information with auditors. The final

plan therefore balances ideal risk assurance coverage with operational realities, making trade-offs as necessary.

Timing	Auditable Unit	Audit Level	Days	Information Access Required	Rationale
Q1	Dangerous capability evaluation (cyber)	Model	100	*Technical:* Evaluation benchmarks, model performance data, uplift measurements. *Procedural:* Evaluation methodology, trigger criteria. *Operational:* Completed evaluations, evaluator interviews.	Highest residual risk; increasing trend; foundational assurance
Q2	Safety classifier effectiveness	Model	100	*Technical:* Classifier architecture, training data, performance metrics, adversarial test results. *Procedural:* Development procedures, monitoring procedures. *Operational:* Production metrics, evasion incident records.	High residual risk; increasing trend; critical control for external misuse
Q3	Incident response capability	System	80	*Structural:* IR policy, team charter. *Procedural:* IR procedures, playbooks. *Operational:* Incident logs, tabletop exercise results, IR team interviews. *Technical:* Containment mechanisms, rollback capabilities.	High residual risk; untested capability

Illustrative audit plan sample (in Appendix A)

What are the main trade-offs?

1. There is a trade-off between how much information auditors can access and how much confidence they can provide, and over which areas. Limiting access can help protect sensitive assets, but it also pushes audits toward high-level reviews, makes it harder to test whether controls actually work in practice, and can leave the highest-risk areas unexamined.
2. Choosing a sourcing model which has a Chief Audit Executive (CAE) internally, such a co-sourcing model, gives flexibility over retaining sensitive information within the security perimeter of the organization without limiting audit scopes. However, an internally sourced model may be perceived as less independent and can take longer to establish than a fully outsourced model.
3. Another trade-off exists around how often audits are run. Infrequent audits reduce the burden on engineering and safety teams, but they risk relying on assurance that quickly becomes outdated as models, systems, and processes evolve.

Conclusion

Ultimately, frontier AI developers will make different trade-offs depending on their risk appetite, organizational maturity, and security posture. Organizations can customize their approaches over time, building toward more comprehensive approaches over time by developing as their models become more capable and risks increase.

A key advantage of internal audit is its ability to fill a critical assurance gap by providing visibility into broader organizational controls and governance, beyond narrow model-level evaluations. It also offers a practical way for companies to meet the requirements of the EU's General-Purpose AI Code of Practice. Developing effective internal audit functions now allows organizations to test and refine these mechanisms before they become mission-critical, while positioning them to meet rising expectations from regulators, investors, and the public.

The goal is not to eliminate all risk or provide absolute certainty, but to establish structured, independent mechanisms that give confidence that catastrophic risks are being actively and responsibly managed. At a time when the decisions of a small number of AI companies may determine how well systemic risks relating to AI are managed, that assurance is increasingly important.