

THE AI INSTITUTE / OPERATING INTELLIGENCE

REGULATORY DECISION BRIEF · 01 OCT / 2026



Ask What the AI Safety Audit Actually Tested

Before relying on an audit, ask what it covered, how hard it looked and whether the failures were fixed.

Boards, C-suite and senior management

Research period: September 2026 – October 2026

First published: 1 October 2026

Current to 1 October 2026

PUBLICATION RECORD

An Institute thesis and decision framework

An AI audit should support a specific operating decision with evidence of its scope, testing effort and remediation—not stand in for a general promise of safety.

WHAT LEADERS SHOULD TAKE FROM THIS PAPER

- | | |
|----|---|
| 01 | Specify the decision the audit must support. |
| 02 | Ask what was excluded and how much testing was attempted. |
| 03 | Separate a voluntary commitment from a legal obligation. |
| 04 | Require evidence that material failures were fixed and retested. |

Research method

Original Institute synthesis of current specialist reporting, a new arXiv preprint, European Commission guidance and Singapore IMDA materials, checked 1 October 2026. The audit-record proposal is Institute analysis. No supplier audit was performed for this article. No private anecdotes or unapproved quotations are used.

First published 1 October 2026 · Updated 1 October 2026 · Research period September 2026 – October 2026 · Version 1.0 · Suggested citation: The AI Institute, *Ask What the AI Safety Audit Actually Tested* (2026).

CONTENTS

Inside this paper

01	Do not approve an adjective The decision
02	A test can miss what it barely looked for Why scope matters
03	Commission a record the business can use The evidence request
04	A pledge does not replace a legal map Where regimes diverge
05	Independence is necessary evidence, not the whole answer Avoid false assurance
06	Make the conclusion expire when its basis changes The next approval
M	Method and limitations
R	References

How to use this brief

Start with the decision and use the framework to challenge the current plan. Figures retain their original populations and definitions. Institute analysis is labelled, and limitations are recorded before the references.

Do not approve an adjective

Before accepting an AI supplier's claim that its system has been independently audited, ask for the decision the audit supports. Permission to run a limited trial, expand a workflow and remove a human checkpoint are different decisions. A reassuring label cannot tell a board which one the evidence justifies.

On 29 September, six AI companies signed a voluntary White House safety accord. Infosecurity Magazine reports commitments to internal controls, internal review, independent external assessment and board oversight. Its 30 September account describes an agreement, not new regulatory requirements.¹ That distinction belongs near the top of any executive briefing about it.

The useful response is to improve what your organisation asks of an audit. Specify the system and operating conditions, the effort spent looking for failures, and the evidence that material findings were addressed. This is the Institute's proposed commissioning discipline, not a claim that the accord itself supplies those details.

A test can miss what it barely looked for

A preprint newly listed on arXiv on 30 September reconstructs selected agent failure behaviours in simplified environments. Its authors found that eliciting different known behaviours required very different amounts of testing. Their automated method targeted behaviours described in advance; it did not solve the harder problem of discovering risks nobody had anticipated. ²

The work is a controlled research exercise, not a census of deployed failures. Its simplified setup and simulated interactions limit what can be inferred about production systems. It nevertheless illustrates a purchasing problem: two reports with no observed failure may represent very different searches.

Ask the evaluator how the test cases were chosen, which conditions were varied and what stopped the search. Was the budget exhausted, a coverage target reached, or an agreed procedure completed? A clean result without that context can be overread. More effort is not automatically better, but the amount and purpose of the effort should be visible.

Keep ordinary task quality separate from the ability to resist misuse or unintended action. Passing one evaluation does not answer every other question. The report should say which property was examined rather than inviting the reader to infer that the whole service is safe.

Commission a record the business can use

Start with a short description of the proposed use. Name the model version, surrounding software, available tools, permissions, data types and relevant languages. Record whether the evaluator tested that configuration or a different one. A supplier’s model-level assessment can be valuable without covering your connected application.

Next, require a method and effort statement. It should describe the test population, number of attempts, repeated runs, human review and exclusions. Where results are sampled, ask how uncertainty is represented. Do not demand an attractive percentage unless its denominator and meaning are clear.

The findings should distinguish an observed failure, a suspected weakness and an untested possibility. Ask for severity in terms of the business consequence, with supporting evidence available through an appropriate confidential channel. A public summary may protect sensitive details; confidentiality should not prevent the accountable decision-maker from understanding the remaining risk.

Finally, request the remediation trail: what changed, who checked the change, which tests were repeated and what remains unresolved. Finding a problem is not the same as closing it. If the proposed fix changes a capability or introduces a new dependency, that trade-off belongs in the approval record too.

EXHIBIT / FOUR PARTS OF A USEFUL AUDIT RECORD

01 Scope The exact system and decision covered. Institute recommendation	02 Effort How the search was designed and bounded. Institute recommendation	03 Findings Observed failures and unresolved uncertainty. Institute recommendation	04 Retest Evidence that material fixes worked. Institute recommendation
--	---	--	---

A pledge does not replace a legal map

The US accord is a voluntary commitment by its signatories. Do not treat it as a new worldwide duty, an exemption from existing law or evidence that a particular product passed an assessment. Nor should an enterprise buyer assume that a frontier-model commitment covers every downstream application it builds.¹

In the European Union, the AI Act imposes additional duties on providers of general-purpose AI models with systemic risk. Commission guidance describes model evaluation, systemic-risk assessment, serious-incident reporting and cybersecurity obligations. It also distinguishes models placed on the market before 2 August 2025, for which the relevant compliance transition runs to 2 August 2027. These are scoped legal duties, not an interchangeable version of the US pledge.³

Singapore's updated Model AI Governance Framework for Agentic AI offers deployment guidance. It is a governance framework, not a universal legal certification.⁴ A company can use such guidance to improve its tests while separately determining what the law requires in each market.

For the UK and other markets, do not copy the EU provider classification into a local compliance checklist without review. For operations across Asia-Pacific, Africa, the Middle East and Latin America, identify applicable sector, privacy and contractual requirements and test the actual languages and infrastructure used. This briefing does not establish equivalent laws or comparable audit capacity across those regions.

Independence is necessary evidence, not the whole answer

An external evaluator can challenge assumptions that an implementation team misses. Ask about relevant expertise, access to the system and any financial or implementation interests. But independence alone does not establish adequate coverage. A technically strong review can still answer a narrower question than the business is asking.

Resist turning audit effort into a competition over the largest number of tests. Repeating easy cases can inflate activity without improving coverage of consequential failures. Conversely, a small, well-designed test may reveal a reason to stop. Ask which decision changed because of the work and what contrary evidence was considered.

Where local language, connectivity or institutional practice differs from the tested setting, record the gap. Use a bounded local trial or additional review to close it before expanding the claim. The absence of comparable evidence is a limit on the conclusion, not evidence that a region is unsafe or that every deployment must wait indefinitely.

Make the conclusion expire when its basis changes

For the next material AI approval, have the business owner and risk lead agree on the decision first. Then commission the evidence needed for it. Attach unresolved findings to the approval, with a named owner and a clear condition for escalation.

Set retest triggers for meaningful changes to the model, tools, permissions or intended use. The old report can remain useful history without being treated as current assurance. Record which changes require a fresh assessment and which can be handled through routine monitoring.

The aim is not another document that says AI was reviewed. It is an approval whose limits are understandable: this configuration, for this use, on this evidence, with these issues still open. That gives a board something it can question—and management something it can act on.

RESEARCH RECORD

Method and limitations

Method

Original Institute synthesis of current specialist reporting, a new arXiv preprint, European Commission guidance and Singapore IMDA materials, checked 1 October 2026. The audit-record proposal is Institute analysis. No supplier audit was performed for this article. No private anecdotes or unapproved quotations are used.

Limitations

The original signed accord could not be retrieved directly; its description is attributed to specialist reporting. The research is a preprint using simplified conditions and known target behaviours, not certification or production failure rates. Legal discussion is selective and does not replace jurisdiction-specific advice. Cross-regional audit outcomes were not established.



Sources and limitations

Source facts link to the original publication. Institute analysis reconciles definitions and turns the material into a management proposition. Frameworks are decision aids, not claims of universal causality. Read every important statistic with its population, date, definition and caveat.

Research cutoff: 1 October 2026. Review cycle: six months, or earlier after a material change in the underlying facts, standards or policy.

REFERENCES

Evidence behind the thesis

01 Infosecurity Magazine: White House safety accord

Alessandro Mascellino, 30 September 2026; describes the 29 September voluntary agreement. Specialist reporting, not the original signed instrument.

02 Slocum et al.: A Reproduction & Lessons for Alignment Testing

v1 submitted 18 September; newly listed 30 September 2026. Preprint. Simplified environments, known target behaviours; not a production incidence study.

03 European Commission: Obligations for general-purpose AI providers

Current official guidance checked 1 October 2026. Distinguishes systemic-risk duties and the transition for models marketed before 2 August 2025.

04 IMDA: Updated Model AI Governance Framework for Agentic AI

20 May 2026 update. Singapore governance guidance, not new law or universal certification; current indexed official description checked.

THE AI INSTITUTE

Make the decision.
Then measure what changes.

Research · Readiness · Capability

theai.institute