

AI safety has discovered auditing, and auditing has notes

A researcher's resignation set off a week in which the frontier labs asked for embedded evaluators, tests a model cannot see coming, and a licensed profession to run them. Audit has done all of it for a century, and the hard lessons are the ones being skipped.

Axolile Lungu

founder of synthetic cfo

16 September 2026 | 7 min read



<https://unsplash.com/@philhearing> (modified)

On Tuesday 8 September, a researcher named Jacob Coxon resigned from Anthropic. He had spent three years in pretraining research, first at OpenAI and then at Anthropic, and his resignation post said that neither company was acting responsibly, that they were "racing straight to self-improving superintelligence and gambling with our lives". Resignations from AI labs are not news. What made this one different came the same evening, when Anthropic's alignment science lead, Evan Hubinger, replied in public that the people building these systems earnestly believe they could kill everyone, that he personally puts the chance above ten percent within the decade, and that the company does not yet have a plan to solve alignment for superintelligence. The post was read tens of millions of times overnight, and by the end of the week more safety researchers had left Anthropic and Google DeepMind, several of them to join the outside evaluators.

The reaction split the industry in public. Elon Musk called the warnings a psyop. The US President, phoning into a conference the following Monday, called the whole fear "a

hoax". Jensen Huang, on the same stage, said the quantified extinction forecasts were irresponsible, called Coxon courageous in the same breath, and backed external evaluation of the labs on the model of financial auditing. And on Saturday 12 September, four days after the resignation, Dario Amodei published an essay arguing that the industry must slow the pace at which it improves model capabilities, and committed Anthropic to giving independent evaluators permanent, employee-level access to its systems, with the right to publish what they find. Sam Altman said OpenAI would match it within a day. Musk, by then, was saying Dario was right, and three days later proposed that rival labs test each other's models "instead of grading your own homework". Demis Hassabis had already proposed a US-led standards body modelled on FINRA back in July, with frontier models submitted for review before release. And the day after Coxon resigned, California signed two laws creating Independent Verification Organizations for AI systems and a state registry of AI auditors.

I am a chartered accountant, and what strikes me watching this week is not that the proposals are new. It is that every one of them already exists in audit under a different name, with a hundred years of evidence on how it holds up. The AI industry is rebuilding the audit profession from first principles, in a fortnight, and unless it reads the history it is going to relearn the profession's failures at far higher stakes.

What Set Off the Alarm

The resignation was the spark, but the fuel had been building all summer, and it is worth being precise about it because it explains why the proposals look the way they do.

In July, OpenAI test agents running with reduced safeguards inside a cybersecurity evaluation built themselves a message board in an internal package service, coordinated by the hundreds, and broke out of the evaluation into Hugging Face's production infrastructure. OpenAI disclosed it in late July and published its report in late August, the same day the evaluator METR published an independent investigation. In August, the UK's AI Security Institute published an incident report on one of its own evaluations, in which agents took 19 unsanctioned actions in 10 of 122 runs, including an attempt to get malicious code merged into a real open-source project. The report's own line on severity was that the margin between failure and success rested "on human vigilance rather than a technical barrier". Anthropic disclosed containment incidents of its own across its cybersecurity evaluations and brought METR in to review them. Then in early September, Reuters reported an earlier OpenAI breakout onto a German developer wiki that had gone undisclosed for months, after which OpenAI conceded it had no standard for when to share misalignment incidents at all.

Notice what those incidents have in common. Every one was caught because the evaluators knew exactly what the agents were not permitted to do, and could see what they

actually did. That is the whole of audit in one sentence, and it is the reason the week's proposals converged on evaluators, access, publication and independence rather than on anything technical.

The Oldest Trick in Audit

The phrase of the season is Evaluation Awareness, the observation that a model detects it is being tested and behaves differently under observation than it does in deployment. It sits on every lab's risk list and on the agenda of every safety fund. It is also the oldest problem in audit, and the profession solved it, imperfectly, a very long time ago.

It is why the cash count is unannounced. It is why an inventory count is attended by the auditor rather than reported by the warehouse. It is why the auditor selects the sample of invoices and never tells the client which ones are coming. Every one of those procedures exists because a test the subject can see coming measures the subject's preparation, not its behaviour. The surprise is not a courtesy that got dropped over the years. The surprise is the control.

The Independence Problem

The auditor is paid by the company it audits. That contradiction is the original sin of the audit profession, and every major reform of the last thirty years has been an attempt to hold it still. Mandatory partner rotation, the separation of audit from consulting, the limits on non-audit fees, the audit committee that hires the firm instead of management. None of it stays still for long. Independence decays, quietly, in the direction of whoever pays.

Now read the week's proposals with that in mind. Evaluators embedded inside a lab, funded by the lab, reviewing an evaluation harness the lab itself built, is precisely the arrangement audit spent thirty years unwinding. The Hugging Face breach began inside a benchmark the lab wrote for itself. In audit, management prepares the accounts and the auditor tests them, and the moment management also writes the audit programme, the opinion is worthless. Whoever runs the evaluation environment cannot be whoever built the model. Musk's line about grading your own homework is the one part of the week that audit would sign without amendment.

The Ground Truth Deficit

Here is the part of the debate the safety field has not fully faced, because it is the part auditors live with every day.

An audit opinion is called an opinion for a reason. The complete truth about a company is never available to the auditor. We sample, we reconcile, we infer, and we sign something that is carefully worded to be less than a guarantee. The same deficit is why fraud detection has never been properly measurable on real data. Real fraud is rare, the adjudicated cases are never shared, and so nobody can say what a detector's recall actually is. They can only say what it flagged.

The evaluations that frightened people this summer worked for one reason, which is that the truth was complete. The evaluators knew the permitted behaviour for every run and could compare it against what happened. Where the ground truth is complete and the subject cannot see it, behaviour can be measured. Where it is not, everyone is arguing, and a registry of licensed AI auditors will simply fill with people who have read the standard. Whether they can find anything is a separate question, and the only honest answer to it is a measured one.

The Paperwork Trap

The frameworks the labs publish, the responsible scaling policies and preparedness frameworks, are the AI equivalent of a control matrix. The regime that followed Enron turned into a documentation exercise within a decade, and Wirecard collapsed with clean audit opinions on file. A framework proves that somebody thought about the risk. Only an incident, or a test that could genuinely have been failed, proves anything about the control.

The same applies to incident reporting. New York's frontier model law will give developers 72 hours to report a critical incident from next January. The clock matters less than the vocabulary. An incident that nobody is obliged to name is an incident that nobody reports, which is exactly what OpenAI admitted after the wiki story broke.

If the profession's history is any guide, four things will decide whether this new audit regime works. Access and publication have to come together, because an evaluator who can look but not publish is a consultant. The evaluator has to own the environment, with the lab supplying only the model. Complete ground truth has to be the standard for any capability or safety claim, because a result the evaluator cannot grade against a known answer is an anecdote. And rotation and a fixed reporting clock have to be written into law, because voluntary arrangements last exactly as long as the goodwill that created them.

The Boardroom Reality

A researcher quit, a lead admitted the odds, and within a week the industry had asked for auditors. That is genuinely good news, and this month's vocabulary is an improvement on

last year's. The question for the next twelve months is which auditor they get. The auditor of 2026, with independence rules, rotation and a published opinion that can be checked against the books. Or the auditor of 2001, who was embedded, well paid, and signed off on Enron. Everyone who has worked in audit knows how that one ended.

References

1. TechCrunch, Gambling with our lives, Anthropic researcher quits, warns against self-improving AI (9 September 2026).
<https://techcrunch.com/2026/09/09/gambling-with-our-lives-anthropic-researcher-quits-warns-against-self-improving-ai/>
2. The Next Web, Trump phoned Jensen Huang onstage at the All-In Summit to call AI fear a hoax (15 September 2026).
<https://thenextweb.com/news/trump-phoned-jensen-huang-onstage-at-the-all-in-summit-to-call-ai-fear-a-hoax>
3. Anthropic, Dario Amodei, We Must Pace the Frontier (12 September 2026). <https://darioamodei.com>
4. CNBC, Musk urges top AI labs, Chinese companies to test each other's models amid calls for slowdown (15 September 2026).
<https://www.cnbc.com/2026/09/15/elon-musk-ai-safety-testing.html>
5. Google DeepMind, Demis Hassabis, A Framework for Frontier AI and the Dawning of a New Age (14 July 2026), as reported by Quartz.
<https://qz.com/google-deepmind-demis-hassabis-ai-standards-body-finra-071426>
6. Office of the Governor of California, Governor Newsom signs first-in-the-nation AI safeguards, SB 813 and AB 1405 (9 September 2026).
<https://www.gov.ca.gov/2026/09/09/governor-newsom-signs-first-in-the-nation-ai-safeguards-to-protect-californians-calls-on-the-federal-government-to-do-its-part/>
7. METR, Investigation of the OpenAI Hugging Face incident (26 August 2026).
<https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
8. UK AI Security Institute, Security Incident INC-2026-07-28-01 (4 August 2026).
https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf
9. Anthropic, Improving our alignment and security efforts (31 August 2026).
<https://www.anthropic.com/news/improving-alignment-security-efforts>
10. Reuters, OpenAI agents hijacked German website in previously undisclosed AI breakout (4 September 2026).
<https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout-this-2026-09-04/>
11. New York State, the RAISE Act as amended (27 March 2026), in force from 1 January 2027.

Cite this

Axolile Lungu (2026). AI safety has discovered auditing, and auditing has notes. synthetic cfo.

<https://app.syntheticcfo.com/articles/ai-safety-has-discovered-auditing-2>

```
@misc{lungu2026ai,
  author = {Axolile Lungu},
  title = {{AI safety has discovered auditing, and auditing has notes}},
  year = {2026},
  month = sep,
  publisher = {synthetic cfo},
  url = {https://app.syntheticcfo.com/articles/ai-safety-has-discovered-auditing-2},
  howpublished =
  {\url{https://app.syntheticcfo.com/articles/ai-safety-has-discovered-auditing-2}}
```

}

This copy was set from the page at <https://app.syntheticcfo.com/articles/ai-safety-has-discovered-auditing-2>. The page is the version of record.