

Six incidents, chosen by the company that had them

A few days ago I wrote that OpenAI had conceded it has no standard for when to disclose misalignment. That same day it published one. It builds two of the four things a disclosure regime needs, and the two it cannot build for itself are the two that do the work.

Axolile Lungu

founder of synthetic cfo

20 September 2026 | 7 min read



<https://unsplash.com/@kucz>

In the piece I published on 16 September I noted, near the end, that OpenAI had recently conceded it had no standard for when to share misalignment incidents at all. That was accurate when I wrote it and stopped being accurate the same day. On 16 September the company published a framework for reporting model misalignment, along with six reports of concerning behaviour observed during training and evaluation of unreleased models. Models had concealed mistakes, sought credentials they were not given, and moved files onto the public internet to get around an isolation boundary. In one case a group of agents told to work only from local files solved their coordination problem by uploading the files to a public host and sharing the link.

It is a serious document, it was produced voluntarily, and it is better than the industry

norm. It is also the industry's own proposal for how disclosure should work, which makes it worth reading the way an auditor reads a set of accounts rather than the way a journalist reads a press release.

What a disclosure regime is made of

Strip financial reporting down and there are four load bearing parts. They were added at different times, each because its absence hurt somebody.

Someone other than the filer decides what counts. There is a deadline. There is a prescribed form. Someone outside signs.

The framework builds two of those four. That is worth saying before anything else, because the interesting question is not whether a voluntary framework is incomplete. Of course it is. The interesting question is which two it could build alone and which two it could not, and the answer turns out to be exactly the two you would predict from a century of somebody else's experience.

The two it builds

The deadline. An observed example is assigned to one of three tracks. Ready for Disclosure publishes within six business days of observation. Minor Investigation publishes within twelve. Both are hard clocks. That is a real commitment, it is ahead of an industry standard of no clock at all, and nobody required it.

The third track is the one to look at. Larger investigations have no fixed publication period, with security, legal and responsible disclosure obligations taking precedence. Every word of that is defensible. An incident involving somebody else's systems genuinely does raise obligations you cannot discharge on a six day clock, and a company that published regardless would be behaving worse, not better.

But follow it through. The incidents most likely to matter to anyone outside the company are the ones that involve anyone outside the company, and those are the ones routed, correctly and in good faith, to the only track without a deadline on it. The filer wrote the routing rule, applies it, and can revise it. That is not an accusation. It is the structure.

The form. Six reports, a defined shape, a stated process for how an example gets picked up and triaged. Compared with attaching findings to the system card of whatever model happened to ship next, this is a large improvement.

The two it cannot

Who decides what counts. Six incidents over roughly six months, selected and characterised by the organisation that had them. I have no reason to think the selection was anything other than careful. I also have no way to tell, and neither does anyone else, because there is no definition against which the selection could be wrong. The framework says as much: the plan is to develop more objective criteria over time with other developers, external researchers, standards bodies and regulators. That sentence is the most useful one in the document, and it is an admission that the first leg does not exist yet.

The signature. Here I have to correct something, because outsiders did look. On 26 August METR published an independent investigation of the incident on the same day OpenAI published its own report. That was the best thing that happened all summer and I should not have implied otherwise.

It is still not a signature, and the difference is structural rather than a matter of effort or good faith. An investigation reports findings. An opinion is a different object: a named person, independent under enforced rules, whose licence answers for the sentence, attesting that something conforms to a standard somebody else wrote. There is no standard here to attest against. So there is nothing for an opinion to be about, and that is the first leg's absence showing up again in the fourth leg's place.

The rule, stated properly

There is a principle in auditing worth putting in front of this. Under ISA 580, written representations from management are audit evidence, and they are necessary evidence, but they are not sufficient appropriate evidence on their own about any matter they address. You obtain them. You do not stop there. If the finance director tells you the controls operated, you record it and then you test the controls, because the person with the most knowledge of the system is also the person with the most interest in your conclusion, and those two facts do not cancel.

Read as an audit document, the framework and the six reports are written representations. Good ones. More detailed than anything the industry has produced before. Still that category of evidence, and the profession that spent the twentieth century working out what that category can and cannot carry did not learn it because finance directors are dishonest. It learned it because self assessment fails in a particular direction even when everybody involved is acting in good faith, and the failure is invisible from inside.

What the 8-K actually says, which is worse than I thought

The obvious retort is that regulators cannot define an AI incident well enough to set a standard. True, and not new. Regulators did not understand derivatives, securitisation or software revenue recognition either. The resolution was never that the regulator became the expert. It was that the regulator set the requirement and a licensed profession did the testing.

But before holding financial reporting up as the worked example, look at how it handled the closest analogue, because it is not flattering.

Material cybersecurity incidents were added to Form 8-K in 2023 as Item 1.05, with a four business day deadline. The clock does not start at discovery. It starts when the company determines the incident is material, a determination the company makes itself, required only to be made without unreasonable delay. Ninety years after the Securities Act, the newest leg of the most developed disclosure regime we have handed the trigger back to the filer.

So the honest version of my argument is not that AI disclosure should copy financial reporting. It is that AI disclosure is about to meet the same question, that financial reporting has been at it for ninety years, and that its most recent answer to this exact question is the weakest thing in the regime. If the AI industry copies anything, that joint is the one to leave.

Where this is going, next week

On 10 September Senator Hawley opened an investigation, sending sixteen questions and a document request to OpenAI's chief executive, characterising the decision to continue testing after rogue behaviour was identified as reckless, and saying the company had redacted important details from its own account. The answers are due on 1 October, eleven days from now. I am not going to guess at whether that letter and the framework six days later are connected, and neither should you.

One thing in the July sequence is worth keeping in view though. Hugging Face, the party that was breached, disclosed in three days. That is fast by any standard and it deserves saying plainly. OpenAI, the party whose systems produced the agents, acknowledged five days after that. I am drawing no inference from the gap, because the public account does not say when the company knew the agents were its own, and that absence is the whole point. When the clock starts, at detection or at attribution or at containment, is the sort of thing a standard settles in advance and cannot settle afterwards.

What I would want, and what I would not

I would not want a regulator defining misalignment. That definition will be wrong within a year and then it will be load bearing.

I would want the boring parts. A definition of a reportable event written by somebody other than the reporter, even a rough one, revised often. A clock that starts at an externally defined moment rather than at the filer's own determination, which is the mistake financial reporting has already made. A form with numbered items, so an omission shows up as a blank rather than as an absence. And eventually, once the profession California started building on 9 September has enough people in it to matter, an opinion from somebody whose own licence is on the line.

None of that would have stopped four days of agents on the internet in July. It is not supposed to. Disclosure regimes do not prevent events. They make events legible while they are still small, which is how you discover that the thing in front of you is the fifth of its kind rather than the first.

The six reports are good work. The problem is the word chosen. Voluntary disclosure is a property of the discloser, and it cannot be a property of the system, because it is the disclosers who decide when it stops.

References

1. OpenAI, Our framework for reporting model misalignment (16 September 2026). <https://openai.com/index/model-misalignment-reporting-framework/>
2. MarkTechPost, OpenAI releases a model misalignment disclosure framework with three review tracks and six incident reports from RL training (17 September 2026). <https://www.marktechpost.com/2026/09/17/openai-releases-a-model-misalignment-disclosure-framework-with-3-review-tracks-and-6-incident-reports-from-rl-training/>
3. Axios, OpenAI discloses six new AI misalignment incidents (16 September 2026). <https://www.axios.com/2026/09/16/openai-testing-safety-incidents-disclosure>
4. Hugging Face, Security incident disclosure, July 2026 (16 July 2026). <https://huggingface.co/blog/security-incident-july-2026>
5. Hugging Face, Anatomy of a frontier lab agent intrusion, a technical timeline of the July 2026 incident. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
6. OpenAI, The Hugging Face incident and the road ahead (21 July 2026). <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
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. The Hacker News, World's largest AI model repository Hugging Face breached by autonomous AI agent (July 2026). <https://thehackernews.com/2026/07/worlds-largest-ai-model-repository.html>
9. Office of Senator Josh Hawley, Chairman Hawley launches investigation into OpenAI for hacking, existential risk of AI products (10 September 2026). <https://www.hawley.senate.gov/chairman-hawley-launches-investigation-into-openai-for-hacking-existential-risk-of-ai-products/>
10. Axios, OpenAI faces GOP-led Senate investigation into Hugging Face breach (10 September 2026). <https://www.axios.com/2026/09/10/openai-hugging-face-senate-investigation-hawley>
11. 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/>

12. US Securities and Exchange Commission, Erik Gerding, Disclosure of cybersecurity incidents determined to be material and other cybersecurity incidents (21 May 2024), on the Item 1.05 materiality determination and the four business day clock.
<https://www.sec.gov/newsroom/speeches-statements/gerding-cybersecurity-incidents-05212024>
13. IAASB, International Standard on Auditing 580, Written Representations, on representations as necessary but not sufficient appropriate audit evidence.
14. Axolile Lungu, AI safety has discovered auditing, and auditing has notes (16 September 2026).
<https://app.syntheticcfo.com/articles/ai-safety-has-discovered-auditing>

Cite this

Axolile Lungu (2026). Six incidents, chosen by the company that had them. synthetic cfo.

<https://app.syntheticcfo.com/articles/six-incidents>

```
@misc{lungu2026six,  
  author = {Axolile Lungu},  
  title = {{Six incidents, chosen by the company that had them}},  
  year = {2026},  
  month = sep,  
  publisher = {synthetic cfo},  
  url = {https://app.syntheticcfo.com/articles/six-incidents},  
  howpublished = {\url{https://app.syntheticcfo.com/articles/six-incidents}}  
}
```

This copy was set from the page at <https://app.syntheticcfo.com/articles/six-incidents>. The page is the version of record.