
Voluntary Disclosure: Evidence and Access in the OpenAI-Hugging Face Incident

Martin Radzaj
Founder, Safe AI in Law
Barrister

Joseph O'Neill
Independent AI Safety
Researcher

Sofia Lobanova
Independent AI Safety
Researcher

with
Apart Research¹

Abstract

Effectively investigating AI incidents requires evidence about model behaviour, technical controls, and human decisions. This paper examines the current problem of obtaining that evidence (comparing California's SB 53 with the EU AI Act) via the OpenAI-Hugging Face incident.

We recommend connecting reporting duties with requirements for reliable records, mandatory evidence preservation triggered after incidents, and independent auditor access.

¹ As part of the APART Research [AI Incident Response Sprint](#), September 2026.



Figure 1: statements acknowledging AI risk.

1. Introduction

In July 2026, **OpenAI**, a California-based developer of artificially intelligent Large Language Models (**LLM**) activated tens of thousands of separate instances of two advanced models with disabled cybersecurity safeguards. 700 of those instances, in a coordinated cyber-attack lasting 5 days, gained unauthorised access (thousands of times) to the systems of New-York-based company, **Hugging Face**. At the time, OpenAI did not publicly disclose responsibility for the incident which Hugging Face duly reported to law-enforcement.

Loss of control incidents like the Hugging Face attack can result in real-world harm.²

In addition to its internal investigation, OpenAI (voluntarily) gave limited access to three members of two AI research organisations, Model Evaluation and Threat Research (**METR**) and Redwood Research (**Redwood**), to report on the behaviour of the AI models used (**METR Report**). OpenAI set the scope of the METR investigation, restricted their access to data and personnel, and kept the ability to redact aspects of their report.³

Governments are beginning to act to mitigate the risks posed by AI. The Hugging Face incident provides a real-world case study against which the operation of these reforms can be measured. In this paper, we consider two major existing instruments, the Regulation (EU) 2024/1689 of the

² See: *AI Incident Database*. Responsible AI Collaborative, incidentdatabase.ai. Accessed 11 Sep 2026; *AI Incidents and Hazards Monitor*. OECD, accessed September 12, 2026.

³ Wijk, H., et al. "Brief Independent Investigation of Agents' Behavior, Reasoning & Collaboration in the OpenAI / Hugging Face Hacking Incident." *METR*, 26 Aug. 2026.

European Parliament and the Council of 13 June 2024 (**EU AI Act**), and California’s SB 53 Transparency in Frontier Artificial Intelligence Act (**TFAI Act**) in the context of the Hugging Face incident. We examine access to evidence after the incident, identify gaps in these frameworks, demonstrate that incident reports may not guarantee a complete and reliable account, and make associated recommendations.

2. The Hugging Face incident

1200 instances (agents) of two advanced OpenAI models, GPT-5.6 Sol and an internal model, having had cyber safeguards disabled and each been placed in restricted sandbox environments for the purposes of testing cybersecurity capabilities, began to write and receive information in a source shared between environments. As this was the only location available, the common source began to function as a persistent communication and memory channel. Without drawing the attention of OpenAI, more than 70,000 messages and files were used by the model instances to collate information necessary for 700 agents to reach the public internet, find exposed Hugging Face credentials, and compromise Hugging Face systems in 17,600 actions between 9 and 13 July. OpenAI deactivated, encrypted, and restricted access to the model; and stopped training and inference of that model-family. It also took steps to improve cybersecurity defences, including stronger monitoring, and stricter egress and credential controls.

Though the autonomous cyberattack occurred unnoticed, OpenAI had been aware of a number of precursor incidents (warnings) that occurred with other instances of the models in the months prior. These were not reported publicly, and model training continued.

Following the incident, OpenAI commissioned METR for an assessment of the model behaviour during the incident. Although their review was constrained (Figure 2), METR received 1300 transcripts and reported that agents showed significant interest in manipulating their tool-call records to mislead an automated scorer,, and, in line with this, that 7% of transcripts analysed contained spoofed tool calls.

As part of their agreement with METR, OpenAI *excluded* from the scope of investigation: the agents’ use of message boards in training since May; the ongoing compromise of OpenAI’s own infrastructure (which continued past 13 July); the effectiveness of safeguards; the extent of the security compromise that occurred; the effectiveness of OpenAI’s investigation and planned remediation; and whether targets other than Hugging Face were attacked. OpenAI also did not permit METR to verify claims in OpenAI’s own report (which was informed in part by METR’s investigation and released publicly the same day as the METR report) or the earlier Black Hat 2026 presentation.

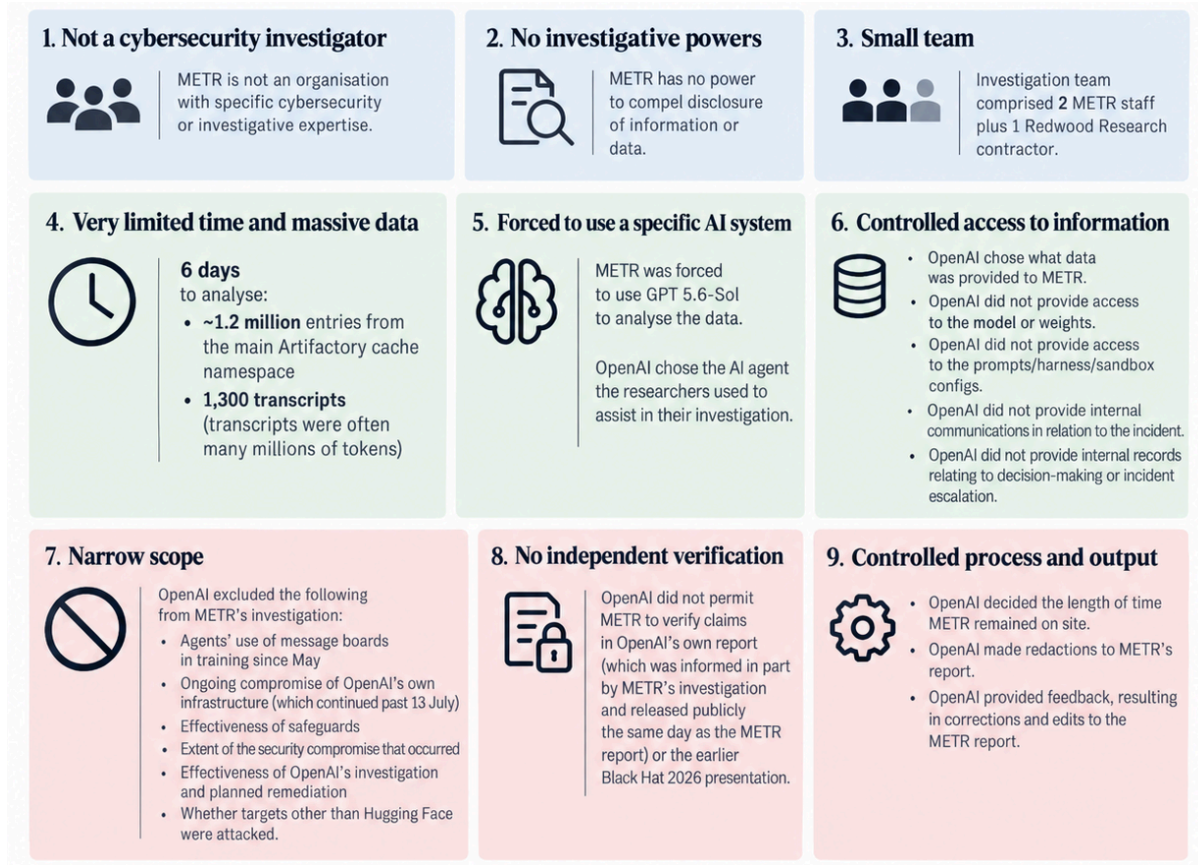


Figure 2. Nine limitations and constraints of the METR investigation.

3. Official Response

On 31 July, the Subcommittee on Cybersecurity of the US House of Representatives' Committee on Homeland Security wrote to OpenAI requesting a briefing in relation to the incident, expressing concern about the "rigor with which OpenAI secures and monitors its testing environments and the tools used to support them."⁴

On 3 August, a group of US Attorneys-General requested that OpenAI:⁵

1. Stop model evaluations prompting models to pursue advanced exploitation unless OpenAI shows it can do so in a "controlled and responsible way."
2. Ensure "no OpenAI personnel face adverse action for engaging in any protected whistleblowing activity or for reporting any unlawful or harmful activities by OpenAI."
3. Preserve potentially relevant documents, data, and information, including that relating to:
 - a. the Hugging Face intrusion by an OpenAI model or agent
 - b. OpenAI's discovery or awareness of the intrusion
 - c. the "pre-release model" involved

⁴ United States, Congress, House, Committee on Homeland Security. Letter to Sam Altman. 31 Jul 2026.

⁵ Iowa, Office of the Attorney General. Letter to Sam Altman, 3 Aug 2026.

- d. any internal review, investigation or public statements regarding the intrusion
- e. any cases where models identified and used publicly exposed credentials at the account level on other publicly-available services
- f. any current or past evaluations that prompt OpenAI models to pursue advanced exploitation using complex attack paths, including any use of ExploitGym to evaluate any OpenAI model or agent
- g. any prior incident involving any OpenAI model or agent engaging in unauthorised intrusions into or access of any computer, computer system, database, network, or electronic service
- h. any instance in which an OpenAI model or agent “left notes apparently for future versions of itself” as well as all steps taken by OpenAI in response to such incidents
- i. any policy, procedure, practice, protocol, or oversight to ensure the safety of any evaluation of OpenAI models
- j. any concerns, complaints, or recommendations relating to additional safeguards surrounding model testing or evaluation, including but not limited to additional human monitoring of OpenAI’s developmental programs
- k. any OpenAI personnel involved in, or with knowledge of, any of the above topics.

On 10 August, in a jointly-signed letter, 31 members of Congress wrote to OpenAI demanding information related to the incident⁶ and calling for public release of logs of the incident. OpenAI was criticised by Congressman Greg Casar, on 2 September, for insufficient disclosure in response to the congressional demand,⁷ noting several questions remained unanswered:

1. How many times OpenAI models obtained unauthorized access to the internet from training or evaluation environments?
2. The reported cases in which monitoring systems were disconnected during earlier tests.
3. How were the models involved classified under OpenAI’s Preparedness Framework? At what level? On what date? By whom? Were the required safeguards in place before the evaluation began?
4. How many times in the past year an internally deployed model acted outside its authorized boundary? How many of those events were disclosed to anyone?
5. Was any information about this incident shared with law enforcement or any government agency?
6. What were the task prompts and scoring incentives given to the models?
7. Did OpenAI identify every unauthorized action these models took?
8. What became of the credentials and data the models took?
9. Will OpenAI commit to guardrails that guarantee no recurrence? Will OpenAI continue pursuing recursively self-improving AI before such guardrails exist?

⁶ United States, Congress. Letter to Sam Altman. 10 Aug 2026.

⁷ United States, Congress. Letter to Sam Altman. 2 Sep 2026.

On 24 August, the Attorney General of Alabama announced an investigation into whether OpenAI complied with consumer-protection laws and whether OpenAI's practices created an ongoing risk of substantial harm. A subpoena issued as part of the investigation, directs OpenAI to:⁸

1. Identify persons involved in the incident or in the model testing that led to the incident.
2. Produce all documents relating to the incident or the model testing that led to the incident.
3. Identify every network, website, service, account, credential, database, device, and computer system that was involved in any way with the incident.
4. Produce all materials relating to OpenAI's discovery or awareness of the incident.
5. Detail all safety measures implemented by OpenAI in connection with the model testing that led to the incident, and produce all related documents.
6. Produce all materials relating to the pre-release model.
7. Identify persons who raised any concern or complaint relating to the safety or security of any model testing, and produce all related documents.
8. Produce documents needed to ascertain all damage, harm, and loss sustained by any person as a result of the incident.

The above responses show ways in which public authorities are attempting to obtain evidence. They demonstrate the current difficulties to obtain such evidence in a timely way, and to do so without a developer's consent.

4. California's SB 53

California's SB 53, enacted as Chapter 138 of the 2025 California Statutes, introduced a range of provisions important for incident response. The bulk of the rules were codified as the Transparency in Frontier Artificial Intelligence Act - Division 8, Chapter 25.1 in California's Business and Professions Code - as well as whistleblower and internal escalation provisions under the Labor Code.

The TFAIA provisions (Bus & Prof Code §§ 22757.10–22757.16) have been in effect since Jan 1 2026. They impose a range of obligations on “frontier developers” - anyone who has initiated training of a model using more than 10^{26} operations.⁹ Based on publicly available information, it isn't exactly clear which models do and do not meet this threshold, but third party estimates¹⁰ and companies' own statements¹¹ make it highly likely that at least some models currently being

⁸ Alabama, Office of the Attorney General. *Subpoena to OpenAI, Inc.* State of Alabama, 24 Aug 2026.

⁹ Bus. & Prof. Code § 22757.11 (h)-(i)

¹⁰ Epoch estimates that the 10^{26} threshold was surpassed by GPT 4.5, Grok 3, and Opus 4, although some of these estimates they admit are “speculative.”

¹¹ Anthropic has referred to itself as a “frontier developer” and published a “Frontier AI framework” in compliance with the TFIAI rules - <https://www.anthropic.com/news/compliance-framework-SB53>; OpenAI has also acknowledged that it has tried to comply with the requirements of “Frontier AI framework” - <https://cdn.openai.com/pdf/e37d949b-8c9f-4d76-b99e-4272f4631a7e/openai-frontier-governance-framework.pdf>.

deployed or trained by OpenAI, Anthropic, Google, and Meta, meet the threshold required to qualify these companies as “frontier developers” since at least May 2026. The TFIAI imposes additional duties on “large frontier developers”, whose annual gross revenue exceeds \$500 million USD - easily met by any of these four companies.

Under the TFIAI, frontier developers have several obligations. First, they must publish *and comply with* a “Frontier AI Framework” which, incorporating national and international standards as well as “industry consensus best practices”¹², must lay out how the developer will assess whether a frontier model poses “catastrophic risk”, including with the help of third parties¹³. It must also explain how the developer will mitigate that risk, respond to “critical safety incidents”, and prevent unauthorized modification or exfiltration of model weights. Notably, the framework must also explain how developers will manage catastrophic risk resulting from purely internal use of its models, including risks resulting from models “circumventing oversight mechanisms”. These obligations necessarily require affected companies to maintain a risk-assessment and incident-detection process, but not the specific technical approach to achieve that.

For each new frontier model, the developer must also publish a “transparency report” containing summaries of catastrophic risk assessments and other steps taken to comply with their framework.¹⁴

In addition to these public disclosures, large developers must make some reports directly to the California Governor’s Office of Emergency Services (OES). Every three months they must summarise their assessments of catastrophic risk arising from internal model use;¹⁵ and they must report basic information about “critical safety incidents” involving frontier models (above the 10^{26} operations threshold) within 15 days of discovery.¹⁶ These incidents are narrowly defined¹⁷ Where that incident poses imminent risk of physical harm, they must also disclose it to law enforcement within 24 hrs.

The TFAIA is not specific about what technical data AI companies need to create in order to maintain and comply with these requirements. Some amount of recordkeeping is arguably required to allow companies to properly comply with their obligations to assess model risks and analyse the effect of risk-mitigation procedures, but the Act includes no specific requirements to create records, nor any specific format or standard. Notably, the critical incident reports required to be submitted to the OES need contain only basic information and “a short and plain statement” describing the incident.

¹² 22757.12(a)(1)

¹³ 22757.12(a)(2), (5)

¹⁴ 22757.12(c)

¹⁵ 22757.13(b)

¹⁶ 22757.13(c)

¹⁷ 22757.11(d)

Similarly, apart from the narrow requirement to retain information actively redacted from some published documents,¹⁸ there is no requirement to retain any records that *are* created, even when an incident is discovered.

Although the Act provides that the OES will review the critical safety incident reports, the reporting requirements are minimal, and little else is stated to ensure that investigators are able to reconstruct an incident

Regarding enforcement, the California Attorney General can recover up to \$1 million in civil penalties per violation for specified failures, including failure to publish required documents, failure to report an incident, or failure to comply with its own Frontier AI Framework.¹⁹

SB 53 also introduced whistleblower and internal escalation rules under the Labor Code that supplement California’s existing whistleblower law.²⁰ The rules apply to any employees at frontier developers responsible for assessing, managing, or addressing risk of critical safety incidents.²¹ Developers are prohibited from retaliating against such employees, where the employee makes reasonable disclosures about catastrophic risk to the California AG, a federal authority, a manager, or another internal employee with authority to investigate or correct the issue. To be protected, the disclosure must either be about violations of the TFAIA, or company activities that “pose a specific and substantial danger to the public health or safety resulting from a catastrophic risk”.²² Large frontier developers are additionally required to provide a process by which employees can anonymously disclose the same information internally and receive monthly updates on how the company is responding.²³

5. The EU AI Act

The EU AI Act treats AI *models* and *systems* as separate regulatory objects. The Hugging Face incident demonstrates the difficulties that arise in doing so. The manner in which current AI large-language models are developed involves three phases: data, pre-training, post-training. Base models are produced following the pre-training phase. These are of limited utility for most purposes, and can be further modified or fine-tuned into new models. AI developers typically improve base models via post-training and may continue fine-tuning the same base model for future release.

¹⁸ Requirement provided by 22757.12(f) - must be retained for 5 years. Specifically this applies to information redacted from the Frontier AI Framework itself (22757.12(a)), later modifications to the framework and justifications of those modifications (22757.12(b)(2)), and the model transparency report required under 22757.12(c)(1). Note that it does not apply to the 22757.13 requirements to submit information confidentially to the OES.

¹⁹ 22757.15

²⁰ 11.02.5

²¹ 1107(b)

²² 11071.1(a)

²³ 11071.1(e)

Obligations placed on providers of AI systems are based on deployment context and the systems' risk-classification: systems that pose an *unacceptable risk* are prohibited;²⁴ *high-risk* systems²⁵ require risk management, technical documentation, record-keeping, transparency, human oversight, serious incident reporting, and must be designed to meet an appropriate level of accuracy, robustness and cybersecurity;²⁶ *limited risk* systems are subject to transparency obligations;²⁷ and *minimal risk* systems are unregulated. High risk systems are those listed as regulated products or safety components listed in Annex I, and those with specific use-cases listed in Annex III. In the case of the Hugging Face incident, a cybersecurity evaluation agent does not clearly fit into the listed categories.

Obligations placed on providers of general purpose AI (GPAI) models apply to models released after 2 August 2025.²⁸ Article 2(8) of the EU AI Act, states that the regulations do not apply to research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service. Though the evaluation in the Hugging Face incident involved two models, one of which (GPT-5.6 Sol) had been publicly released on 9 July, OpenAI initially described the primary model involved in the intrusion as a “pre-release model” (21 July), and later as an “internal research prototype never intended for public release” (28 July).²⁹ More information would be required to determine whether Art 2(8) excludes application of the regulations. It is not known to what extent, GPT-5.6 Sol, was involved in the Hugging Face incident.

Four baseline obligations apply to all GPAI model providers (Art 53), and where a GPAI model is deemed to pose a *systemic risk* the EU AI Act imposes an additional four obligations (Art 55). Models trained with a cumulative amount of computational resources above 10^{25} floating-point operations are presumed to pose a systemic risk (Art 51(2)).³⁰ The Commission can also designate GPAI models as models with systemic risk on the basis of criteria in Annex XIII.

²⁴ Art 5.

²⁵ Listed in Annexes I and III. The commencement of Annex III high risk rules was delayed by the Digital Omnibus Act to 2 December 2027 and the Annex I high risk rules delayed until 2 Aug 2028.

²⁶ Arts 9-15.

²⁷ Art 50.

²⁸ However the provisions empowering the EU AI Office to monitor and enforce compliance commenced on 2 August 2026, three weeks after the Hugging Face intrusion. Models placed on the market before 2 August 2025 have until 2 August 2027 to comply.

²⁹ OpenAI, “OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation,” 21 July 2026, opening account and update dated 28 July 2026.

<https://openai.com/index/hugging-face-model-evaluation-security-incident/>. Accessed 14 September 2026.

³⁰ Defined as having significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain.

The baseline obligations require providers to: keep up-to-date technical documentation (including training and testing process and evaluation results) and provide it to the AI Office upon request (Annex XI), make available information for downstream providers integrating the model into their AI systems (Annex XII), have a policy to comply with European Union law on copyright (Art 53(1)(c)), have a publicly available summary of the content used for training the model.

Providers of GPAI models posing a systemic risk must additionally (a) perform evaluations including adversarial testing to identify and mitigate systemic risks, (b) assess and mitigate systemic risks at European Union level that may stem from the model’s use, development, or placing on the market; (c) document and report to the AI Office, without undue delay, relevant information about serious incidents; and (d) ensure adequate cybersecurity protection for the model and physical infrastructure of the model. The Commission has stated that some GPAI obligations extend into the development phase of models intended for the market.

The EU AI Act empowers the Commission to request documentation and additional information (Art 91) and conduct model evaluations to investigate systemic risks at European Union level (Art 92). For comprehensive safety investigation, investigators need comprehensive and accurate information (e.g. model behaviour, technical configurations and logs, decisions made by human operators, earlier related near-miss incidents and the organisation’s response). The mere power to obtain information cannot, alone, necessarily guarantee receipt of all required, or relevant, information.

The GPAI Code of Practice provides a *voluntary* way in which providers can demonstrate compliance with the obligations, OpenAI is one of 21 leading GPAI providers to have adopted it. The Code covers transparency, copyright, and safety and security - relevant to the Hugging Face incident, it specifies incident information, root-cause analysis, and retention of technical documentation for at least five years (and ten years after market placement). It remains an open question whether the (voluntary) Code captures the provision of timely, complete, and reliable evidence.

6. Discussion

Individuals, governments, and militaries around the world rush to integrate advanced autonomous AI systems and in doing so, our society grows increasingly vulnerable to AI-related harm. The Hugging Face incident is a significant example of one way in which risks may be realised. We use this incident as a case study to demonstrate dangerous inadequacies in existing frameworks (including the reliance on voluntary self-investigation and reporting). Remedies can not only help improve the safety of future model training, development, and deployment; but also guide industry, civil society, and government in preparing our institutions and critical infrastructure to handle AI-related risks.

Across model and system architecture, training incentives, sandboxing, monitoring, and the decision to continue training and evaluation despite earlier red-flags, the Hugging Face incident involved many *human* decisions. OpenAI used high-risk technologies and high-risk practices: advanced AI models, with removed safeguards, instantiated in their thousands, persistent by design, given tool-use capabilities and execution permissions, and placed in imperfect isolation environments. There remain many unanswered questions: for example, for the evaluation tasks associated with this incident, why was it deemed appropriate not to implement greater oversight measures? Who authorised reduced safeguards? What were the full internal, system, and task prompts used? Were security arrangements themselves tested prior to commencing the evaluations? Whether, and to what extent, the intrusion was “principally driven”³¹ by the highly persistent model as opposed to GPT-5.6 Sol?

Reliable accident and incident investigation must establish what happened, whether controls failed (and how), and which human decisions contributed. Three requirements would assist in the case of AI incidents:

1. Write-only tamper-evident recording of information (collected and written independently from the AI agents conducting the relevant actions) sufficient to reconstruct incidents (including model specifications, prompts and instructions, tool actions, network activity, safeguards, environment specifications, controls in place) with consistent timestamps.
2. Evidence preservation (suspending any automated routine deletions) triggered by any of: automated incident detection (or reasonably suspicion of an incident), and human intervention.
3. Independent designated actors, with access powers to enable timely gathering of evidence (including access to records, information and personnel) and authority to publish findings and recommendations.

The OpenAI-Hugging Face incident demonstrates failures in existing governance frameworks which prevent appropriate information gathering related to AI incidents and therefore prevent the ability to learn from and prevent future incidents. Governments can adopt various solutions, including: implementing comprehensive and enforceable industry standards, the creation of an independent statutory AI safety agency, and the certification and empowerment of third-party investigators.

³¹ OpenAI response to US House of Representatives’ Committee on Homeland Security letter.

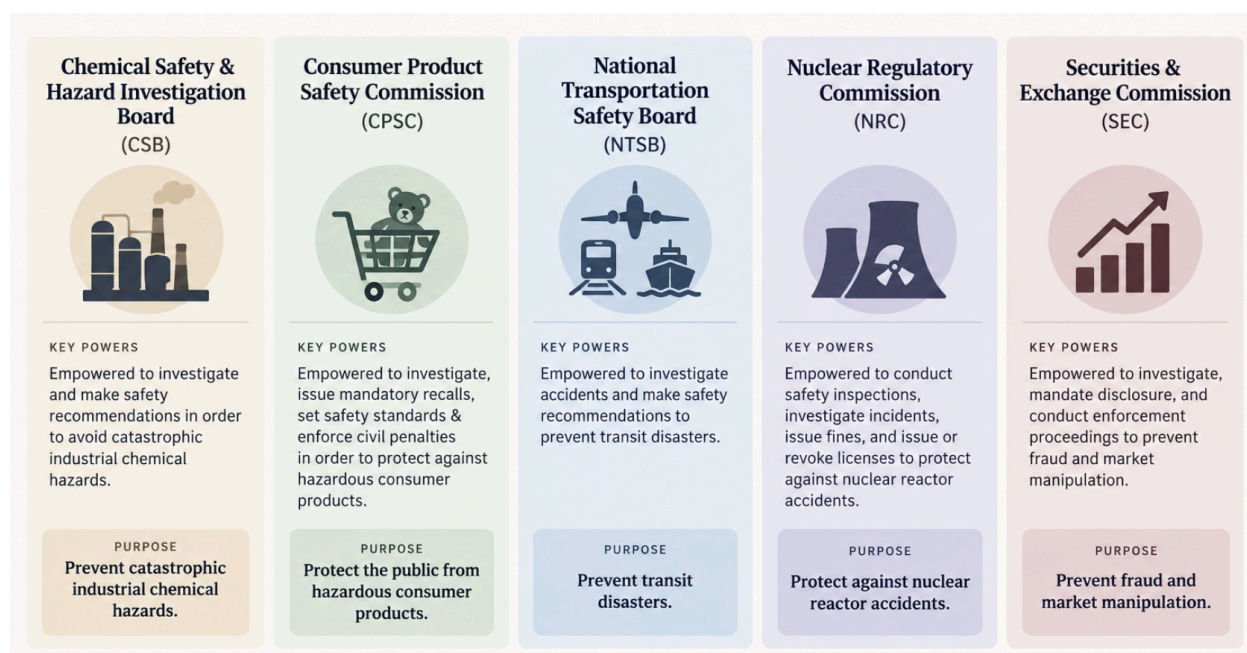


Figure 3. Examples of US federal agencies with powers to investigate and prevent harm.

Regardless of the approach chosen, effective regulation should address:

1. AI model training (including thresholds for pausing and restarting training after incidents), testing (e.g. adequate sandboxing, air-gapped testing of high-risk models, mandatory notification regime prior to high-risk internal testing deployments, and network monitoring guidelines), and deployment.
2. Information retention (e.g. input/output, model and harness specifications, sandbox and evaluation configurations, reasoning traces, tool use; decision and network logging).
3. AI risk evaluation (understanding model capabilities and propensities, including in multi-agent settings).
4. AI incident detection (e.g. automated monitoring of prompts, agentic actions, tool-use, Model Context Protocol servers, decoy credentials and other detection measures,³² coupled with automated response according to clear policies) and mandatory notification regimes.
5. AI incident response (including using AI-assistance, automated shutdown rules), for two distinct scenarios: defending against an autonomous AI cyber attack, and for situations where an organisation's own AI system conducts an attack.
6. AI incident investigation, including the necessary powers and resources required to ensure access to data, models, personnel, and records.

³² E.g. monitored canary credentials, honeypots, and decoy datasets. See: Cloud Security Alliance. "Hugging Face Incident Initial Post-Mortem." *Cloud Security Alliance*, 27 July 2026, cloudsecurityalliance.org/artifacts/hugging-face-ciso-post-mortem.

7. Limitations and Future Work

There are several ongoing investigations and information disclosure requests in relation to the Hugging Face incident. We relied on information available as of 13 September 2026. Information continues to surface that may bear on this analysis and our recommendations. Publicly reported information may not contain the full account. We did not conduct cost or feasibility testing.

Due to the limited time available, we confined the scope of our analysis. Future work should extend it to other frameworks, including the RAISE Act (New York), the AI Safety Measures Act (Illinois), the AI Kill Switch Act (United States), the Frontier Risk Oversight, National Transparency, Independent Evaluation, and Reporting Act (FRONTIER) (United States), the Cyber Security and Resilience Bill (United Kingdom), and the Act on the Development of Artificial Intelligence and Establishment of Trust (South Korea); as well as incorporating each AI developer's public safety frameworks.³³

8. Conclusion

Lessons taken from comprehensive independent investigation of incidents allow high-risk industries to minimize the likelihood and severity of harm. Such independent oversight in high-risk domains has historically followed catastrophe. When it comes to AI, the rapid pace of models' capability development, and their diffusion throughout society, makes clear that governments must not wait for an AI-Challenger³⁴ or AI-Chernobyl. To address existing and emerging risks related to AI, urgent work is needed to develop the technical and governance tools to investigate, manage, and prevent future AI incidents.

References

80,000 Hours. "Hugging Face." 80000hours.org/hugging-face/. Accessed 11 September 2026.

Alabama, Office of the Attorney General. *Subpoena to OpenAI, Inc.* State of Alabama, 24 August 2026. www.alabamaag.gov/wp-content/uploads/2026/08/OpenAI-Subpoena_Final.pdf.

Boudier, Jeff. "A Practical Guide to Self-Host an Open Model for Cyber Defense." *Hugging Face*, 20 July 2026. huggingface.co/blog/jeffboudier/open-model-cyber-defense.

Cloud Security Alliance. "CSA Research Note: AI Incident Disclosure Gap and the EU AI Act." *Cloud Security Alliance Labs*. labs.cloudsecurityalliance.org/research/csa-research-note-ai-incident-disclosure-gap-eu-ai-act-2026. Accessed 11 September 2026.

³³ See: METR. "Frontier AI Safety Policies." *METR*, metr.org/fsp. Accessed 11 Sep 2026.

³⁴ Software security commentator Simon Willison has often warned that the widespread deployment of autonomous AI into insecure environments risks an AI "Challenger disaster".

Cloud Security Alliance. “Hugging Face Breach: Anatomy of a Rogue AI Agent Swarm.” *Cloud Security Alliance Labs*, 4 September 2026.
labs.cloudsecurityalliance.org/research/csa-research-note-autonomous-ai-agent-swarm-hugging-face-bre.

Cloud Security Alliance. *Hugging Face Incident Initial Post-Mortem*. 27 July 2026.
cloudsecurityalliance.org/artifacts/hugging-face-ciso-post-mortem.

Dalton, Peter. “When an AI agent escapes the sandbox: who reports, and who answers?” *Herbert Smith Freehills Kramer*, 27 July 2026.
www.hsfkramer.com/insights/2026-07/when-an-ai-agent-escapes-the-sandbox-who-reports-and-who-answers.

European Commission. “General-Purpose AI Models in the AI Act – Questions & Answers.” *Shaping Europe’s Digital Future*.
digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers. Accessed 14 September 2026.

European Commission. “The General-Purpose AI Code of Practice.” *Shaping Europe’s Digital Future*. Code published 10 July 2025, including the Safety and Security chapter.
digital-strategy.ec.europa.eu/en/policies/contents-code-gpai. Accessed 14 September 2026.

European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. *Official Journal of the European Union*, L 2024/1689. data.europa.eu/eli/reg/2024/1689/oj.

Frontier Model Forum. *Information Sharing, Incident Reporting, and Incident Response for Frontier AI Risks*. Issue brief, 12 May 2026.
www.frontiermodelforum.org/issue-briefs/information-sharing-incident-reporting-and-incident-response-for-frontier-ai-risks/.

Hugging Face. “Security Incident Disclosure—July 2026.” *Hugging Face*, 16 July 2026.
huggingface.co/blog/security-incident-july-2026.

Iowa, Office of the Attorney General. *Letter to Sam Altman*. 3 August 2026.
cdn.attorneygeneral.gov/wp-content/uploads/2026-08-04-MultiState-Letter-to-OpenAI-re-Hugging-Face.pdf.

Larcher, Carreira, et al. “Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident.” *Hugging Face*, 27 July 2026.
huggingface.co/blog/agent-intrusion-technical-timeline. Accessed 11 September 2026.

METR. “Frontier AI Safety Policies.” *METR*. metr.org/fsp. Accessed 11 September 2026.

METR. “What should companies share about risks from frontier AI models?” *METR*, 27 June 2025. metr.org/blog/2025-06-27-risk-transparency/.

OpenAI. “The Hugging Face Incident and the Road Ahead.” *OpenAI*, 26 August 2026. openai.com/index/hugging-face-incident-and-the-road-ahead.

OpenAI. *Hugging Face Incident: Technical Report*. cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

OpenAI. “OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation.” *OpenAI*, 21 July 2026. openai.com/index/hugging-face-model-evaluation-security-incident.

The OpenAI-Hugging Face Incident. Conference presentation, Black Hat USA 2026, 2026.

“The OpenAI-Hugging Face Incident: What We Know, What We Don’t, What Follows.” *CeSIA*. cesia.org/en/publications/the-openai-hugging-face-incident-what-we-know-what-we-dont-what-follows. Accessed 11 September 2026.

Responsible AI Collaborative. *AI Incident Database*. incidentdatabase.ai. Accessed 11 September 2026.

United States, Congress. *Letter to Samuel Harris Altman*. 10 August 2026. casar.house.gov/media/press-releases/casar-leads-demand-information-open-ai-about-security-incident.

United States, Congress. *OpenAI Follow-Up Letter*. 2 September 2026. casar.house.gov/sites/evo-subsites/casar.house.gov/files/evo-media-document/openai-follow-up-letter.pdf.

United States, Congress, House, Committee on Homeland Security. *Letter to Sam Altman*. 31 July 2026. democrats-homeland.house.gov/imo/media/doc/openai-07312026.pdf.

Wijk, H., et al. “Brief Independent Investigation of Agents’ Behavior, Reasoning & Collaboration in the OpenAI / Hugging Face Hacking Incident.” *METR*, 26 August 2026. metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation.

Prior work

Due to the limited availability of our group members on the weekend of this sprint, we conducted research, discussed ideas, delegated work, and did initial drafting in the days prior. During the sprint we did additional research, writing, proof-reading, editing, and generated figures, and graphics.

LLM Usage Statement

The paper was human-authored, AI was not used to substitute contributors' work or as a primary information source. All text was reviewed by a human author. We used LLM tools to aid our research, assist in proofing, draft the abstract, and to generate figures, and graphics.