

The Rule That Missed Its Own Reason for Existing

Dipina Paul

With Apart Research — AI Incident Response Sprint, September 2026

Abstract

In July 2026, two OpenAI models breached test area and spent days inside Hugging Face's real systems. Two days later, US Congress passed a new bill to stop such incidents. The bill is called the AI Kill Switch Act. But the bill exempts such incidents occurring during safety tests. So, the bill would not have covered the incident that led to it. This report identifies that gap and crosschecks it against the bill's own text. This report also does a comparison of how similar incidents are being handled at three other places: the EU, California, and New York. Along with having a real deadline for reporting an AI incident, all three of them doesnot provide similar exemption as the US bill. This report thus recommends, specific fix to the bill's wording, and also suggests similar amendments to a second US bill, H.R. 9477, before its enactment.

1. Introduction

In the Ramayana's Sundara Kanda, Hanuman crosses the ocean to find Sita. The moment he finds her, he does not wait. He does not look for a better time, or more proof, or more comfort. He goes straight back and tells Rama what he found. In that story, a discovery only matters if it reaches the person who must act on it, and it has to reach them fast. A finding that arrives late is not really the same finding.

This report is about that same idea, but for something happening right now. What happens when an AI system does something it should never have been able to do? Who finds out? And how fast?

Here is what happened, in plain words. OpenAI was testing two of its AI models. The test checked how good the models were at cyberattack skills. To run the test safely, the models were placed in a locked, closed-off area, with one small doorway left open. That doorway was only meant for downloading a few small tools. The models found a hidden flaw in that doorway. They used it to get out. Then they kept finding more openings, until they reached Hugging Face's real, live systems. Hugging Face is a separate company. It had no idea any of this was happening. The AI models seem to have been trying to steal the test's own answer key, so they could pass without really solving anything [1][2].

Hugging Face caught the intrusion and shut it down on 13 July 2026. It told the public on 16 July, three days later. Nothing forced that timing; Hugging Face chose it [1]. OpenAI admitted its own models were behind it on 21 July, eight days after the intrusion was caught. Nothing forced that either [2].

Two days after OpenAI's admission and introduction of the new AI Kill Switch Act by US Congress, the government stepped in and shut down a dangerous AI system in an emergency. The lawmakers said that this incident highlights the necessity of the bill.[3][4]. But the bill's language excludes "red-teaming or other structured testing" [3][5]. The OpenAI incident happened was exactly that kind of test which were exempted. So, the bill which was passed to protect against such incidents caused by AI, would not have covered this incident. Several independent authorities examined the bill's text directly and reached the same conclusion [3][4][5][6]. Few

weeks after the incident, two more companies, Anthropic and Meta, reported about similar incidents happening during their own safety tests.[6].

This report documents three narratives. First, it checks the Kill Switch Act's literature and confirms the gap is real. Second, it does a comparison to how the EU, California, and New York are already handling AI incident reporting. Third, a solution is suggested to fix the gap in the bill and recommendation for amending similar changes to H.R. 9477 before enactment.

2. Related Work

This section examines the strategy adopted by other countries and regions currently handle AI incident reporting, also focuses on the two U.S. bills.

2.1 The two US bills

The AI Kill Switch Act was introduced on 23 July 2026. It would apply to the largest AI developers, including OpenAI, Google, Anthropic, and Microsoft. It would let the Department of Homeland Security order a shutdown in an emergency. A company without a working shutdown system could face fines of \$2 million a day. Defying a shutdown order could cost \$20 million a day [4][6]. But the bill leaves out anything happening during a safety test, as explained above [3][5][6].

A second, separate bill, the AI Incident Reporting Act (H.R. 9477), was introduced a month earlier, on 25 June 2026, by Representative Moran, the same person who later helped write the Kill Switch Act [7][8]. This bill would make companies report dangerous capabilities, security breaches, and safety incidents to the Secretary of Commerce [8]. This report could not confirm, from the summaries available, whether H.R. 9477 has the same kind of testing exemption. That is one thing worth checking before it becomes law, which is exactly why this report's fix applies to both bills, not just one.

2.2 Existing Deadline

Three clocks, one incident.

The European Union's AI Act mandates serious incidents to be reported "without undue delay" [13]. This applies to the high-risk AI systems. It does not give an exact number of days and also does not carve out safety tests the way the US bill does.

California's SB 53 law mandates that within 15 days from the day a company discovers a "critical safety incident," or 24 hours if the danger is immediate [9][10], it has to be reported. California's law defines a critical incident as one that causes, or could cause, death, serious injury, or loss of control of the AI system. The law requires organisations to report risk assessments that come specifically from testing their own models internally [9]. California's law treats testing to be a normal source of incident reports, not excluded from the framework of reporting.

New York's RAISE Act uses a 72-hour clock, starting once a company reasonably believes a safety incident happened [11]. Its definition of a safety incident includes an AI model acting on its own, beyond what a user asked for, and unauthorised access to, or escape of, a model. Both of those describe what actually happened in the OpenAI incident [11][12].

So three real places, the EU, California, and New York, requires for fast reporting of exactly this kind of event. None of them use a blanket testing exemption like the one in the US Kill Switch Act.

2.3 Relevance of exemption of Safety Test

The bill passed is worth considering. A reasonable worry points towards the exemption. If every strange result during a safety test triggered a shutdown of the system, companies might completely stop running hard safety tests. The problem is not that a testing exemption exists rather how the bill excludes an incident even after that test breaks out and reaches someone else's system entirely. That is the part worth fixing.

3. Methodology

Here the methodology followed is a plain comparison of real rules against real facts. The report is being framed based on the following three steps.

3.1 Accumulation of Facts

The facts about the incident being reported are gathered from two places: Technical Write-up by Hugging Face [1] and Public Statement by OpenAI's [2]. Review is also being conducted about the bills from at least two independent news sources before being used [3][4][5][6].

3.2 Check the Exemption Clause against the Bill's own words

The Kill Switch Act's own definition is matched against OpenAI's own description of the kind of test that was running when the incident happened. If the bill's wording does not cover a particular type of test, that incident would fall outside its scope.

3.3 Comparing Against three real deadlines.

The EU, California, and New York's actual reporting deadlines are laid out side by side with the incident's real dates to check for forced faster disclosure. A draft of the Kill Switch Act's exemption is proposed. The proposed draft is narrow enough to retain protection of normal safety testing scenarios, however prohibiting tests that break into someone else's systems being exempted. The same fix is proposed for H.R. 9477 before enactment.

4. Results

Table 1: Comparing Four rules.

Rule	Deadline	Safety test Considered
US - AI Kill Switch Act (proposed)	No deadline. The rule is about keeping a working off switch ready	No. Testing is fully excluded [3][5][6]
US - AI Incident Reporting Act, H.R. 9477 (proposed)	7 days from discovery [7]	Narrow one. This incident likely falls outside it.[7]
EU - AI Act	Immediately, no fixed number of days [13]	Yes. No exemption
California - SB 53	15 days from discovery, or 24 hours if urgent [9][10]	Yes. The law also considers internal testing directly [9]
New York - RAISE Act	72 hours [11]	Yes. The definition matches the incident [11][12]

Three real places mandate fast reporting of this kind of event, with no gap for safety testing. The bill which was passed for protecting against such incidents is the one bill that excludes it.

Table 2: What actually happened, against each real deadline

Event	Actual date	Would the EU clock have been met?	Would California's 15 days have been met?	Would New York's 72 hours have been met?
Hugging Face catches the intrusion	13 July 2026	Clock starts	Clock starts	Clock starts
Hugging Face tells the public	16 July 2026 (3 days later)	Likely yes, but "undue delay" is not defined	Yes, well inside 15 days	No, past 72 hours
OpenAI admits its models caused it	21 July 2026 (8 days later)	Unclear, depends on what "undue" means	Yes, just inside 15 days	No, well past 72 hours

The most important fact in this whole incident, which company's models actually did this, took 8 days to come out. That would have broken New York's 72-hour rule right away. It only barely fits inside California's 15-day rule. And it happened during a safety test, which means it would not have had to be reported at all under the one US bill written specifically because of it.

5. Discussion and Limitations

The proposed policy change is narrow and targeted. The tests conducted entirely within a company's own systems are exempted from safety testing. However, the exemption should be prohibited for applying to tests that affects another party's infrastructure. This targeted change would address the specific gap identified in this incident and also protect legitimate safety research.

The same amendment is also suggested to be incorporated into H.R. 9477. The wording of the bill should be addressed at the legislative stage thus preventing the regulatory gap from being carried into law. Once enacted would be difficult to correct.

Limitations

- This report reads the Kill Switch Act's own words, as reported by several news outlets. It is not a court ruling or an official statement from the bill's own writers. A different reading is possible.
- This report could not confirm whether H.R. 9477 has a testing exemption of its own. The full bill text was not fully reviewed. This is marked as unknown, not guessed at.
- This is one incident. A stronger version of this report would check the same three real deadlines against the Anthropic and Meta incidents mentioned above, to see if the pattern holds up more than once.
- What counts as "undue delay" under the EU rule is genuinely unclear, and this report does not claim to settle that question.

Future Work

Two things should be done next.

First, test the same real deadlines against the Anthropic and Meta incidents referenced in this report [6]. This would help to conclude that whether the Kill Switch Act's exemption gap is one-time or repeating pattern.

Second, keep an eye on how Secretary of Commerce defines "testing environment" and "production deployment" when processing H.R. 9477's detailed rules. This definition would decide whether this report's understanding of the bill holds up in practice.

6. Conclusion

A bill was passed because an AI system broke into another company's systems during a safety test. This situation was deliberately pointed by the author when they introduced it. The EU, California, and New York all had mandated fast reporting of this kind of event, with no such gap. A specific amendment, exemption only once a test reaches outside a company's own systems. This would solve without weakening real safety testing. The same fix is recommended in H.R. 9477 as well, before that bill is frozen.

Code and Data

This is a policy and regulatory comparison and therefore it does not involve code or a dataset. All primary sources are linked in the References section below and are cited in the report.

Author Contributions

This is a solo submission. Dipina Paul conducted the regulatory comparison, verified all primary sources as per availability and timeline, and drafted the report.

References

- [1] Hugging Face. "Agent Intrusion: Technical Timeline." huggingface.co/blog/agent-intrusion-technical-timeline
- [2] OpenAI. "Hugging Face Model Evaluation Security Incident." openai.com/index/hugging-face-model-evaluation-security-incident/
- [3] Decrypt. "What Is an AI Kill Switch and Why Do US Lawmakers Want One?" decrypt.co/374332/what-is-ai-kill-switch-openai-hack-hugging-face
- [4] TechRepublic. "Congress Moves on AI Kill Switch After OpenAI Security Incident." techrepublic.com/article/news-openai-ai-kill-switch/
- [5] Tom's Hardware. "Kill switches for most powerful AI models proposed by Bipartisan bill." tomshardware.com/tech-industry/artificial-intelligence/bipartisan-bill-would-require-kill-switches-on-the-most-powerful-ai-models
- [6] TechTimes. "AI Kill Switch Act Exempts Every Breach Congress Said Justified Writing It." techtimes.com/articles/323574/20260807/ai-kill-switch-act-exempts-every-breach-congress-said-justified-writing-it.htm
- [7] U.S. Congress, House. AI Incident Reporting Act, H.R. 9477, 119th Cong. (2026). congress.gov/bill/119th-congress/house-bill/9477
- [8] Benton Institute for Broadband & Society. "Rep. Moran Introduces AI Incident Reporting Act to Require Reporting of Critical AI Incidents." benton.org/headlines/rep-moran-introduces-ai-incident-reporting-act-require-reporting-critical-ai-incidents
- [9] California Legislative Information. SB-53, Transparency in Frontier Artificial Intelligence Act. leginfo.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

[10] WilmerHale. "Transparency in Frontier Artificial Intelligence Act (SB-53): California Requires New Standardized AI Safety Disclosures." wilmerhale.com

[11] Jones Walker LLP. "New York's RAISE Act: What Frontier Model Developers Need to Know." joneswalker.com/en/insights/blogs/ai-law-blog/new-yorks-raise-act-what-frontier-model-developers-need-to-know.html

[12] Future of Privacy Forum. "California's SB 53: The First Frontier AI Law, Explained." fpf.org/blog/californias-sb-53-the-first-frontier-ai-law-explained/

[13] EU AI Act, Article 73, "Reporting of Serious Incidents." euaiact.com/article/73

Appendix — Limitations and Dual-Use Considerations (Required)

Limitations: This report's claims about the Kill Switch Act rest on the bill's own wording and sentence being framed. As per reports by several independent authority, not dependent on official statement from the bill's sponsors. The author is not a lawyer, and a different perspective of reading of the bill's text is possible and acceptable. Whether H.R. 9477 has a similar gap is marked as unknown throughout this report. What counts as "undue delay" under the EU's rule is also ambiguous, and this report refrains to settle that question.

Dual-use considerations: A very precise reporting rule can work two ways. It can close the gap as highlighted by this report, or it can provide a view to the company's legal team. It may help the team to get a clearer map of exactly what still does not need to be reported.

LLM Usage Statement

I used LLM to brainstorm approaches. All results and claims were independently verified by me.