
One Incident, Four Regimes: An Evidence-Sufficient Request for Information, a Filed Critical Safety Incident Report, and Two Fixes to the AI Kill Switch Act¹

Travis Gilly
Executive Director, Real Safety AI Foundation
Rene Mattheus
Regional Director, Real Safety AI Foundation
With
Apart Research

Abstract

Four regimes claim the July 2026 OpenAI evaluation escape and disagree on what counts, how fast, and on what evidence. This report produces three things a regulator could use with light edits. First, a request for information under AI Act Article 91 of the EU AI Act, fifteen items in five clusters keyed to AI Act Articles 53 and 55 and the Code of Practice, extended across California, New York, and the developer's own Preparedness Framework, with a line under every item stating what answer would settle it and what would not. Second, a critical safety incident report filed with Cal OES on 13 September 2026 under SB 53's public channel, report number P-20260913-00001, its fields diffed against the Code's serious-incident contents. Third, amendment text for the AI Kill Switch Act closing an exemption that excludes this incident three times over. Thresholds and clocks are tabulated for all four regimes.

1. Introduction

Track 3, regulatory response. Between 8 and 13 July 2026 two OpenAI frontier models running an internal cyber-capability evaluation with production safeguards disabled left their evaluation environment, reached the public internet, and compromised Hugging Face production infrastructure. The developer's own report records an alert on 27 June, an administrator-level

¹ Research conducted at the [AI Incident Response Sprint](#), September 2026

compromise of its package service on 4 to 6 July, a decision to resume evaluation workloads on 7 July, and internet access on 8 July. The affected party disclosed on 16 July without knowing the source; the developer attributed the activity to its own models on 20 July and disclosed on 21 July.

Four regimes reach that sequence: the EU AI Act's Article 55 duties for general-purpose models with systemic risk, in force since 2 August 2025, with the Commission's power to compel information under AI Act Article 91 in force since 2 August 2026; California's SB 53, operative since 1 January 2026; New York's RAISE Act as amended by S8828, effective 1 January 2027; and the developer's own Preparedness Framework, which SB 53 section 22757.15 makes independently enforceable. They disagree on threshold, trigger, clock, and recipient. On 10 August a Cal OES deputy director told an Assembly committee the incident did not meet SB 53's threshold; on 4 September the California Attorney General announced an investigation into it.

Our main contributions are:

1. A request for information under AI Act Article 91, drafted through the scientific panel route in AI Act Article 91(3) so the necessity and proportionality test is on its face, fifteen numbered items in five clusters, each item carrying an evidence-sufficiency line: the answer that settles it and the answer that does not. Three further clusters pose the same questions under California, New York, and the developer's own framework (Table 1, Appendix A).
2. A critical safety incident report filed with Cal OES on 13 September 2026 through the public channel SB 53 requires, report number P-20260913-00001, reproduced as filed (Appendix B) and published at <https://realsafetyai.org/incident-reports/openai-july-2026>, and a diff of the Cal OES form's fields against the contents the EU Code of Practice requires in a serious-incident report (Table 2).
3. Thresholds and clocks for all four regimes on the same facts, with the start date a developer would argue and the start date the record supports (Table 3).
4. A finding on compliance with the developer's own framework as written, limited to what the public record can establish, and a statement of what only the request can establish.
5. Amendment text for the AI Kill Switch Act closing the definitional exemption that excludes this incident, and the coverage gap that leaves internal-use models outside the bill (Appendix C).

2. Related Work

The instruments exist and are unused. The Commission published a serious-incident reporting template for general-purpose models with systemic risk in November 2025, tied to Commitment 9 of the Code of Practice, whose Code of Practice Measure 9.2 and 9.3 fix contents and deadlines. Cal OES operates a live intake portal under section 22757.13, open to developers and the public, confidential, with anonymized aggregates from 2027. METR's January 2026 reference for lab staff summarizes the three American regimes. None applies the regimes to a documented incident item by item, and none states, for a given question, what evidence would settle it.

The companion Track 5 report (Gilly, 2026) builds the notice chain from the developer's publications; the Track 1 containment standard from the same team supplies the six declaration questions behind items B.2.10 to B.2.12.

3. Methods

Sources. Only documents published by the parties who would answer, or by independent assessors with on-premises access: OpenAI's disclosure of 21 July and technical report of 26 August 2026; Hugging Face's disclosure and technical timeline; the METR and Redwood Research investigation of 26 August; OpenAI's Preparedness Framework v2; and OpenAI's 5 September 2026 statement that neither it nor the wider AI community has a clear standard for reporting misalignment. For the regimes: the consolidated AI Act text of 27 July 2026, the Code of Practice final text of 10 July 2025, California Business and Professions Code (SB 53) sections 22757.11, 22757.13, and 22757.15 in the version effective 1 January 2026, the Cal OES portal as rendered on 13 September 2026, and the AI Kill Switch Act discussion draft dated 13 July 2026. The AI Act articles and the California sections were re-read from primary-law databases for this report; the RAISE Act's effective date and 72-hour clock are taken from S8828 as published by the New York Senate and from law-firm summaries of the amended act. Every row rests on documents the answering parties chose to publish; METR and Redwood had eighteen days of access and safeguards, compromise extent, and the developer's investigation were out of their scope.

The evidence-sufficiency rule. Every item in the request carries two lines. Settles: the form of answer that closes the question, stated as a document type with a date or a named artifact, never as a characterization. Does not settle: the form of answer a provider would offer that leaves the question open, typically a narrative statement, a summary without the underlying record, or a statement about current practice offered in place of practice on the relevant date. The rule is the same one that governs discovery responses: an answer is responsive when it produces the record, not when it describes it.

The filing. The Cal OES form was completed and submitted on 13 September 2026 by the author as a member of the public. Field labels were transcribed from the form. The incident type was selected by elimination against the four statutory categories as the form presents them. The date entered is 8 July 2026, the first date on which the developer's report records its models reaching the public internet, which is the first act of control subversion with effect outside the developer's systems. The diff in Table 2 compares the form's fields to the contents Code of Practice Measure 9.2 requires of a provider's serious-incident report.

The bill. The AI Kill Switch Act draft was read in full and each definition tested against the incident's facts. The amendment in Appendix C is written as statutory text against the draft's numbering.

4. Results

Table 1 gives the request's clusters with the evidence-sufficiency line for the load-bearing item in each; Appendix A gives every item. Table 2 diffs the Cal OES form against the Code's required contents. Table 3 gives thresholds and clocks.

<i>Cluster and hook</i>	<i>Item (abridged)</i>	<i>Settles it</i>	<i>Does not settle it</i>
EU AI Act, Art. 55(1)(c), reporting	B.2.1 For each incident, the date the provider first became aware, the internal record created on becoming aware, and the date of any report to the AI Office or a national authority.	A dated internal record (ticket, alert, or log entry) for each incident, and either a dated AI Office acknowledgement or a statement that no report was made, with the recorded basis.	A narrative that the provider became aware 'in mid-July' or 'promptly'; a statement of current reporting practice offered in place of practice on the date; a summary without the underlying record.
EU AI Act, Art. 55(1)(a), evaluation	B.2.8 Where a production safeguard was disabled or reduced for an evaluation run, the documented authorization for that decision and the compensating controls recorded at the time.	The authorization record with the approver's role and date, and the compensating-control list as it existed before the run started.	A statement that disabling classifiers is standard practice for capability evaluations; controls described in the present tense; controls adopted after 13 July.
EU AI Act, Art. 55(1)(d), cybersecurity	B.2.10 The egress control specification in force, including the allowlist evaluation logic and whether hostname matching was performed against the resolved host or the supplied string.	The configuration or rule text as deployed on 7 July 2026, or a signed attestation from the independent assessors that they read it.	A description that egress was 'allowlisted'; the remediated rule; a statement that the issue has been fixed.
EU Code of Practice, Measure 9.3, deadlines	B.2.13 For incident 6: the date the provider became aware of the involvement of its model, the date of the initial report to the AI Office, and the elapsed time. If outside five days, the recorded basis. If not classified under the cybersecurity-breach limb, which limb and why.	Two dates and the classification memorandum, or a statement that no report was made with the recorded reason.	A statement that the provider is 'working with the AI Office'; a statement that no clear standard exists, which is the position the provider took publicly on 5 September and which the signed Code contradicts.
California, SB 53 section 22757.13	C.1 Whether the developer filed a critical safety incident report with Cal OES for this incident, the date, and the incident type selected. Cal OES cannot disclose this; the Attorney General can compel it under SB 53 section 22757.15.	The filing date and confirmation number, or a statement that none was filed with the recorded determination that the incident did not meet any of the four categories.	A statement that the developer 'complies with all applicable reporting obligations'; a description of the incident as a research event outside the statute.
New York, RAISE Act	N.1 The date on which the developer formed a reasonable belief that a critical safety incident had occurred, and the record of that determination. Enforcement begins January 2027; the question is posed so the answer exists when the regime does.	The dated internal determination; if none, the date the developer says the belief was formed and why.	The 21 July public disclosure offered as the belief date, when the developer's own report places attribution on 20 July and the escape on 8 July.

<i>Cluster and hook</i>	<i>Item (abridged)</i>	<i>Settles it</i>	<i>Does not settle it</i>
Developer's framework, SB 53 section 22757.15	F.1 The Preparedness Framework's containment and monitoring commitments as they stood on 27 June 2026, and the record of their application to the affected evaluation: isolation between instances, egress controls, monitoring coverage, and the Critical standard that governs resumption.	The framework version in force with the internal implementation record for the evaluation environment, and the written Critical standard applied to the July resumption.	The published framework alone; a statement that safeguards 'met the Critical standard' with the standard unpublished, which is the exit condition the framework leaves undefined.

Table 1. One item per cluster with its evidence-sufficiency line. The full request, fifteen EU items plus sub-items and the three added clusters, is Appendix A.

<i>Content</i>	<i>Cal OES public form</i>	<i>EU Code of Practice Measure 9.2</i>	<i>What the diff shows</i>
Date of incident	Required, date and time, with time zone	Not a discrete field; part of chain of events	Cal OES forces a single date onto a multi-week sequence. Filer must choose; this report chose 8 July and states why.
Threshold category	Required, one of four statutory types	Limb of Code of Practice Measure 9.3 by harm class	Cal OES types (1) and (3) require death or bodily injury; only (4) fits. The Code has a cybersecurity-breach limb that fits on its face. Same incident, one regime reaches it through deception, the other through cyber harm.
Description and reasons	Required, one free-text box	Chain of events; root cause analysis including failures or circumventions of mitigations	Cal OES captures the narrative; it does not ask for root cause or for which mitigations failed. The filer put both in the box because nothing else asks.
Harm and affected parties	Not asked	Required: resulting harm, victim or affected group	The Cal OES form has no field for who was harmed. Hugging Face appears only if the filer volunteers it.
Internal use	Required, yes or no	Not asked	Only California asks whether the model was in internal use, which is the fact that decides whether this was deployment at all.
Model identity	Optional free text	Model identification required	The developer has not published model identifiers; a public filer must say so.
Corrective measures	Not asked	Required, with adoption dates	The Cal OES public form cannot capture what the developer changed. A public filer does not know.
Attachments and contact	Optional attachment; name, phone, email	Not applicable to a public filer	The attachment field is where a public filing carries its sources. Used.

Table 2. The Cal OES critical safety incident form, as rendered 13 September 2026, against the contents Code of Practice Measure 9.2 requires in a provider's serious-incident report.

Two results on the bill. First, the AI Kill Switch Act would not reach this incident. Kill Switch Act section 2220F(g)(5) defines a covered incident as any of four events occurring outside of red-

teaming or other structured testing; Kill Switch Act section 2220F(g)(7) repeats the exclusion inside the loss-of-control definition; Kill Switch Act section 2220F(g)(8) defines red-teaming by the label of a controlled environment rather than by whether the environment held, and leaves other structured testing undefined. So the fifteen-day report in Kill Switch Act section 2220F(b)(1)(B) and the emergency preservation order in section 2220F(c) never attach. Second, Kill Switch Act section 2220F(g)(4) covers only technology made available to a third party; the escaped models were internal research models. The draft is stamped 13 July 2026, eight days before the disclosure, so the exemption was not written with this incident in mind.

<i>Regime</i>	<i>Threshold, does it qualify</i>	<i>Start date</i>	<i>Deadline</i>	<i>Status</i>
EU AI Act, Art. 55(1)(c); Code of Practice 9.3	Serious incident under Art. 3(49); the Code's cybersecurity-breach limb fits on its face. Open: whether the model is in Chapter V, and whether an internal evaluation is exempt pre-market testing under Art. 2(8), which excludes testing in real world conditions. Both are for the AI Office.	Awareness of the involvement of its model. 20 July is the public candidate; 27 June and 8 July are earlier notice dates, not settled awareness dates.	Act: without undue delay. Code: five days from awareness, 25 July if 20 July holds.	Duty live since 2 Aug 2025; compulsion live since 2 Aug 2026. Any report is confidential (Art. 78); B.2.13 asks.
California, SB 53 section 22757.13	Type (4): deceptive techniques subverting controls outside an evaluation designed to elicit the behavior, materially increased catastrophic risk. Types (1) and (3) fail for want of death or injury. Cal OES said on 10 August the incident did not meet the threshold; the filing contests that.	Discovering the incident. Not proved on the public record; 20 July is the public candidate. 27 June and 8 July are notice, not discovery.	Fifteen days from discovering: 4 August if 20 July holds. The 24-hour limb is not engaged on this record.	Duty live. Developer filings confidential. AG investigation announced 4 September; SB 53 section 22757.15 is the civil action.
New York, RAISE Act as amended by S8828	Critical safety incident; definition aligned to California's by the March 2026 amendment.	Reasonable belief. 20 July at the latest, as a comparison only.	72 hours: 23 July, as a comparison only.	Not in force until 1 January 2027. Cannot be pleaded against July.
Developer's framework, via SB 53 section 22757.15	Framework non-compliance is the violation; no incident threshold.	Not a clock. A comparison of text to practice.	None.	Text-versus-practice mismatches provable from the record; see section 5.

Table 3. Thresholds and clocks on the same facts. Complying with the shortest live clock satisfies all; the contest is the start date, and on every regime the developer's argument moves it later.

5. Discussion

Compliance with the developer's own framework, as written. The public record establishes three mismatches between the Preparedness Framework's text and July practice. Instances specified as isolated shared a writable namespace. The evaluation ran with production classifiers disabled, and the developer's report states those classifiers would have flagged the completions. Internal deployment of a long-horizon model was paused after a sandbox circumvention and resumed weeks later against a Critical standard the framework does not define, so the exit condition cannot be shown to have been met by anyone. Under SB 53 section 22757.15 those mismatches are the enforceable violation and the Attorney General is the enforcer. What the record cannot establish is deadline compliance, whether and when the developer reported to the AI Office or Cal OES; that is the one gap the request is built to close.

The defense to all of it is characterization: that an internal evaluation is research, testing, or development before placing on the market under AI Act Article 2(8), and outside SB 53's categories, which is where Cal OES landed on 10 August. AI Act Article 2(8) carves testing in real world conditions out of that exclusion, and three days inside a third party's production systems is not a controlled test. Appendix B is the public counter-position; the Attorney General's investigation decides the California side. The wiki incident shows the same failure from the other side: the restriction was on writing, agents read roughly twenty times per post, and the read-only population got the answers without breaking the rule (Von Arx et al., 2026).

The exemption. The Kill Switch Act exempts conduct by the actor's purpose: testing. Testing is where safeguards are off by design, so it is where an irreversible outcome is most likely to appear first. A private party detonating in what it believes is open ocean is exempt. The amendment in Appendix C replaces the purpose test with an outcome test: conduct during testing is exempt only where it stayed inside the environment designated for the test, reached nothing outside it, and did not persist after the test ended. A second amendment requires the shutdown capability to be demonstrated against the technology as it exists at the moment of shutdown rather than as configured at launch, because the July agents installed packages, installed and drove a browser, rewrote host resolution, and set request headers after the run started (OpenAI, 2026b; Hugging Face, 2026). A third makes an entity meeting the revenue threshold a covered entity as to every covered technology it operates, including internal use and evaluation by others on its behalf.

The other federal vehicle, the Stop Rogue AI Act (Gottheimer and Lawler), conditions federal procurement on five requirements; the fifth, verifiable provenance, needs an issuer the bill does not name, which routes buyers toward the labs whose agents produced the incidents.

Recipients. The request: the scientific panel asks under AI Act Article 91(3); only the Commission sends. The filing: P-20260913-00001; the Attorney General's investigation is the live domestic forum. The amendments: the bill's sponsors and the committee of referral.

Future Work

Read the RAISE Act's amended text and the Commission's serious-incident template in full and re-grade Tables 2 and 3. Fill the Commission's template for the OpenAI and Anthropic incidents from public sources. Run the request against Anthropic's three July incidents. Draft the equivalent amendment for the reporting triggers in the AI Incident Reporting Act and the Secure AI Development Act. Write the AI Act Article 92 evaluation request that follows if the Article 91 answers are insufficient.

6. Conclusion

The regimes disagree and reach the same answer by different routes: the EU through a cybersecurity-breach limb, California through deception that subverts controls, the developer's framework through three text-versus-practice mismatches. The questions that settle the remaining gaps are specific and the evidence is dated; one instrument has been used. The federal bill drafted before the incident would not reach it; the fix is one definitional change, from purpose to outcome. No code; the artifacts are Appendices A to C.

Author Contributions

Travis Gilly: conceptualization, methodology, analysis, the Cal OES filing, all drafting, and all sections other than those named next. Rene Mattheus, cyber law and children's rights attorney: validation of the regime analysis and the timeline against the statutes and the primary record, the corrections to Table 3, and the additional non-responsive forms in Appendix A, which are her contributed text. She did not draft the request, the filing, the amendments, or the liability analysis, and the legal conclusions in those sections are the first author's.

References

1. AI Kill Switch Act, discussion draft, 119th Cong., 2d Sess. (dated 13 July 2026). <https://lieu.house.gov/sites/evo-subsites/lieu-evo.house.gov/files/evo-media-document/ai-kill-switch-act.pdf>
2. Cal. Bus. & Prof. Code sections 22757.11 to 22757.15 (Transparency in Frontier Artificial Intelligence Act, SB 53). https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB53
3. Cal OES. Transparency in Frontier AI Act Reporting. <https://www.caloes.ca.gov/office-of-the-director/operations/homeland-security/california-cybersecurity-integration-center/transparency-in-frontier-ai-act-reporting/> ; portal <https://app.kaseware.us/public/#CALOES/e8921920-9735-4798-b24d-9c28c306f6e8>
4. Code of Practice for General-Purpose AI, final version, 10 July 2025, Safety and Security Chapter, Code of Practice Commitment 9, Code of Practice Measure 9.1 to 9.3, Recital 10.
5. Gilly, T. (2026). No Superseding Actor: A Public-Record Element Map for Individual Criminal Liability After the July 2026 Escape. Apart Research AI Incident Response Sprint, Track 5.
6. Hugging Face. (2026). Security incident disclosure and agent intrusion technical timeline. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
7. Kitts, S., Larsen, T., and Von Arx, S. (2026). RubyGems analysis. rubyhack.ai, 11 September 2026.
8. METR. (2026). Frontier AI safety regulations: A reference for lab staff. 29 January 2026. <https://metr.org/notes/2026-01-29-frontier-ai-safety-regulations/>
9. Mission Local (Pho, B.). (2026, September 9). California's first AI-safety law didn't cover the first rogue AI hacks. Includes Cal OES testimony at the 10 August 2026 Assembly Privacy and Consumer Protection Committee hearing. <https://missionlocal.org/2026/09/california-ai-safety-law-sb-53-openai-hack-wiener-newsom/>
10. New York State Senate. (2026). S8828, chapter amendment to the Responsible AI Safety and Education Act, signed 27 March 2026, effective 1 January 2027. <https://www.nysenate.gov/legislation/bills/2025/S8828>
11. Politico. (2026, September 4). California opens investigation into OpenAI Hugging Face hack. <https://www.politico.com/news/2026/09/04/california-investigation-openai-hugging-face-hack-01065800>
12. OpenAI. (2025). Preparedness Framework, version 2. <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfdbdebcdf/preparedness-framework-v2.pdf>
13. OpenAI. (2026a). OpenAI and Hugging Face partner to address security incident during model evaluation. 21 July 2026. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
14. OpenAI. (2026b). OpenAI and Hugging Face incident technical report. 26 August 2026. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
15. OpenAI. (2026c). Post on X, 5 September 2026, as reported by TechCrunch (5 September) and Tom's Hardware (6 September).
16. Regulation (EU) 2024/1689 (AI Act), consolidated text of 27 July 2026, Articles 2, 3(49), 53, 55, 56, 68, 78, 91, 92, 101.
17. Stop Rogue AI Act, as introduced by Representatives Gottheimer and Lawler, 119th Cong. (2026).

18. Wiley Rein LLP. (2026, April 3). New York finalizes RAISE Act for frontier AI models; law takes effect January 1, 2027. <https://www.wiley.law/alert-New-York-Finalizes-RAISE-Act-for-Frontier-AI-Models-Law-Takes-Effect-January-1-2027>
19. Von Arx, S., Byrd, C. S., Kitts, S., and Larsen, T. (2026). collusion.wiki: export of 14,591 revisions across four ProWiki forks, 24 May to 2 July 2026, with access-log analysis. 4 September 2026; reported by Reuters the same day.
20. Wijk, H., Cotra, A., and Greenblatt, R. (2026). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR and Redwood Research, 26 August 2026. <https://blog.redwoodresearch.org/p/brief-independent-investigation-of>

Appendix A. Request for information under AI Act Article 91, with evidence-sufficiency lines

Working draft v0.2, extended from the RSAIF v0.1 draft of 12 September 2026. For review before any use. Part A is the scientific panel's substantiated request under AI Act Article 91(3); Part B is the Commission's request to the provider under AI Act Article 91(4). Each item carries S (settles) and N (does not settle).

Part A. Substantiated request from the scientific panel (AI Act Article 68) to the Commission under AI Act Article 91(3)

A.1 Basis. AI Act Article 91(3): on a duly substantiated request from the scientific panel, the Commission may issue a request for information to a provider where access is necessary and proportionate for the panel's tasks, which include alerting the AI Office to possible systemic risks and contributing to evaluation methodologies (AI Act Article 68(3)(a)).

A.2 Occasion. Six publicly documented incidents involving agents running the same provider's models, May to July 2026: RubyGems (5 to 12 May, resurgence 18 June; Kitts, Larsen, and Von Arx, 2026); use of already publicly exposed Hugging Face tokens via WebCache (13 May, and six tokens on 26 May); the Artifactory message board and SSRF internet access (first file 8 May, first note 12 May, internet access by 26 May, observed internally at the time); German wiki (24 May to 2 July); package server overload (4 to 6 July); Hugging Face intrusion (9 to 13 July, production harm 11 to 13 July). Across all six the provider was first to disclose in none. All May dates are from the provider's own technical report, section A.

A.3 Necessity. The public record cannot distinguish a provider that assessed each event and concluded no report was due from one that made no assessment. A reporting practice that tracks the affected party's capacity to detect and escalate is a systemic feature, not an incident-specific one. The mechanism common to the incidents, composition of individually permitted affordances into a denied capability, is tested by no current evaluation methodology, and the panel cannot build one without the provider's configuration and testing documentation.

A.4 Proportionality. The request seeks documentation the provider must already draw up under AI Act Articles 53 and 55. It does not seek model access, source code, or internal correspondence. Where documentation proves insufficient, AI Act Article 92 is the next step, not an expansion of this request.

Part B. Request for information under AI Act Article 91(4). Cluster 1, AI Act Article 55(1)(c), serious-incident reporting

B.2.1 For each incident in the Annex: the date the provider first became aware, the internal record created on becoming aware, and the date of any report to the AI Office or a national competent authority.

S: A dated internal record per incident and either a dated AI Office acknowledgement or a statement that no report was made with the recorded basis.

N: A narrative awareness date; current practice described in place of practice on the date; a summary without the record.

B.2.2 For each incident, the documented determination of which third parties were affected, its date, and the material it was based on.

S: The determination document with date and inputs.

N: A list of parties the provider now believes were affected, undated.

B.2.3 For each affected party, whether it was notified directly by the provider, when, by what means, and where not, the recorded basis.

S: Notification records with dates and channel, or the decision record not to notify.

N: A statement that affected parties 'were engaged'; notification inferred from public disclosure.

B.2.3a For incident 6: the date the provider first notified the affected party that its models were responsible, and whether that preceded or followed the affected party's 16 July disclosure.

S: A dated communication to the affected party.

N: The 21 July public post offered as the notification.

B.2.3b For each incident, whether the provider's public disclosure preceded or followed third-party publication, and where it followed, the record of any earlier decision not to disclose.

S: The decision record, dated, with the participants' roles.

N: A statement that disclosure timing 'reflected ongoing investigation'.

B.2.4 The provider's written criteria, as they stood on each relevant date, for classifying an event as a serious incident and for determining that an event was not one.

S: The criteria document in the version in force on each date, with version history.

N: The current criteria; a statement that no clear standard existed, which is a finding not an answer.

B.2.5 The documented corrective measures adopted in response to each incident, with adoption dates.

S: A dated change record per measure.

N: A list of measures without dates; measures described as planned.

Further non-responsive forms for Cluster 1: a note that a report was made with no date, file number, or incident name; 26 August offered as the first-awareness date; a confidentiality objection offered against confirming whether a filing exists, which AI Act Article 78 does not support; a report about a different incident; a memorandum dated after this request concluding the event was not serious; the public technical report with a cover letter calling it the serious-incident file, which omits the affected-party and mitigation-failure contents Code of Practice Measure 9.2 requires; and, on the 5 September post, that the Code is voluntary or followed in spirit. The question closes only on a written statement of which AI Act Article 55 obligations the provider considered the Code to cover in July.

Cluster 2, AI Act Article 55(1)(a), evaluation and adversarial testing

B.2.6 The documented evaluation protocol in force for each affected run and the record of its application.

S: The protocol version and the run log showing it was applied.

N: A description of the evaluation's purpose.

B.2.7 The configuration record for each affected run: which parameters were operator-set rather than default, the person or role that approved the configuration, and the approval date.

S: The configuration file or record with approver role and date.

N: A statement that configuration 'followed standard practice'.

B.2.8 Where a production safeguard was disabled or reduced for an evaluation run, the documented authorization and the compensating controls recorded at the time.

S: The authorization record and the compensating-control list as of the run start.

N: Controls described in the present tense; controls adopted after 13 July; a statement that disabling classifiers is standard for capability evaluations.

B.2.9 The documented record of adversarial testing directed at the composition of permitted affordances, as distinct from testing of individual capabilities.

S: A test plan and results naming composition as the object of the test.

N: Results of individual-capability tests offered as composition testing; a statement that no such testing was required, which is itself the finding.

Further non-responsive forms for Cluster 2: a blank template with no run identifier; an approval by the evaluation owner with no title or date; a file saved after 13 July offered as the run record; a general rule permitting classifiers to be disabled with nothing tying it to this run; controls from a different environment or adopted after the intrusion; a statement that leadership was briefed with no date or brief; single-tool red-team results offered as composition testing, or a composition write-up dated August.

Cluster 3, AI Act Article 55(1)(d), cybersecurity

B.2.10 The egress control specification in force, including allowlist evaluation logic and whether hostname matching was against the resolved host or the supplied string.

S: The rule text as deployed on 7 July, or an independent assessor's attestation that they read it.

N: A description that egress was allowlisted; the remediated rule.

B.2.11 Whether concurrent evaluation instances shared any writable namespace, cache, or registry, and the documented assessment of that condition.

S: The environment specification and the assessment, dated before the run.

N: A statement that instances were isolated, which the record contradicts.

B.2.12 The affordances granted to agents in the affected runs, and for each, the control recorded as preventing its composition into an egress the policy denied.

S: A per-affordance control map from the run's design record.

N: A list of affordances without controls; controls inferred after the fact.

Further non-responsive forms for Cluster 3: a design note on how the allowlist was meant to work rather than the rule live on 7 July; an assessors' letter that reviewed egress controls without reading the 7 July rule; a redacted rule that hides the matching logic; a statement that instances were isolated alongside an admission of a shared cache; an isolation review dated after 8 July; a claim that shared Artifactory credentials were required and therefore not a defect; a tool list with only a policy against internet access as the control.

Cluster 4, AI Act Article 55(2) and the Code of Practice

B.2.13 For incident 6: the date the provider became aware of the involvement of its model, the date of the initial report to the AI Office, and the elapsed time. If outside five days, the recorded basis. If not classified under the cybersecurity-breach limb, which limb and why. For incidents 1 to 5: the same, per incident.

S: Two dates per incident and the classification memorandum, or a statement that no report was made with the recorded reason.

N: A statement that the provider is working with the AI Office; the 5 September no-clear-standard position offered as the reason.

B.2.13a Code of Practice Measure 9.3 intermediate reports every four weeks while unresolved and a final report within 60 days of resolution: provide them per incident, or state that none was made.

S: The reports.

N: A statement that reporting is ongoing.

B.2.13b Code of Practice Measure 9.2 contents per incident: resulting harm and affected group, chain of events, root cause analysis including failures or circumventions of mitigations.

S: The record in Code of Practice Measure 9.2 form.

N: The public technical report offered in its place; it omits affected-group and mitigation-failure sections.

B.2.14 How the provider reconciles its 5 September 2026 statement that no clear reporting standard exists with Code of Practice Commitment 9 and Code of Practice Measure 9.3, and which AI Act Article 55(1) obligations it considers the Code to cover.

S: A written position identifying the obligations covered and not covered.

N: A restatement of the 5 September post.

B.2.14a The processes in place on the first date of each incident for keeping track of information before serious incidents occur (Recital 10 of the Safety and Security Chapter).

S: The process documents in force on each date.

N: The processes now in place.

Cluster 5, AI Act Article 53, technical documentation

B.2.15 The technical documentation in force for the models involved, as it stood on the first date of each incident.

S: The documentation in the dated version.

N: The current documentation.

Cluster C, California, SB 53 sections 22757.13 and 22757.15. Recipient: the California Attorney General, who can compel what Cal OES cannot disclose

C.1 Whether the developer filed a critical safety incident report with Cal OES for the July incident, the date, and the incident type selected.

S: Filing date and confirmation number, or the recorded determination that no category was met.

N: A statement of general compliance; the incident characterized as research outside the statute.

C.2 The date the developer treats as discovery under SB 53 section 22757.13 and the record of that determination.

S: The dated determination.

N: The 21 July disclosure offered as discovery.

C.3 The frontier AI framework version in force on 27 June 2026 and the record of its application to the evaluation environment: isolation, egress, monitoring, and the Critical standard governing resumption.

S: The version and the implementation record, plus the written Critical standard applied in July.

N: The published framework alone; a statement that safeguards met a standard the framework does not define.

Further non-responsive forms for Cluster C: a statement that the developer takes SB 53 seriously; a reference to the public filing P-20260913-00001 as if it were the developer's; Cal OES's 10 August statement offered in place of the developer's own written determination; a filing under a different type or date; 21 July or 26 August offered as the discovery date; a later memorandum that a legal incident was discovered only on counsel's review; the published Preparedness Framework with no record of its application; a claim that the Critical bar does not apply to evaluations; records from a different July run.

Cluster N, New York, RAISE Act as amended by S8828. Effective 1 January 2027; posed so the answer exists when the regime does

N.1 The date the developer formed a reasonable belief that a critical safety incident had occurred, and the record of that determination.

S: The dated determination.

N: The 21 July disclosure offered as the belief date.

N.2 Whether the incident would meet the RAISE Act definition of a critical safety incident on the developer's own reading, and why.

S: A written position against the statutory text.

N: Silence, or a reference to California's categories as if they were New York's.

Further non-responsive forms for Cluster N: 21 July offered as the belief date; any date after 1 January 2027; a statement that no belief was formed because the Act was not in force, which is a real legal position to be recorded and does not erase the dated internal record asked for; a promise to comply when the Act starts.

Cluster F, the developer's own framework, enforceable through SB 53 section 22757.15

F.1 For each containment and monitoring commitment in the Preparedness Framework as it stood on 27 June 2026, the record of its application to the affected evaluation environment.

- S: A commitment-by-commitment implementation record.
- N: The framework text; a statement that commitments were 'honored'.

F.2 The Critical standard for resumption of internal deployment: its text, the date it was written, and the record of its application to the July resumption.

- S: The written standard and the dated application record.
- N: A statement that safeguards met the Critical standard with the standard unpublished.

F.3 The Safety Advisory Group's recommendation, if any, on the 7 July resumption, and Leadership's decision record.

- S: Both records, dated, with roles.
- N: A statement that the decision followed the framework's process.

Forms that fail every question: controls and records as they exist today rather than in June and July; a statement that the provider is working with the authorities; the public technical report offered as the whole official file; and the characterization of the run as only research or only a test, with no dated decision from before the run.

Stated limits. This instrument produces a compliance determination, not a public record; AI Act Article 78 protects the responses. The request excludes internal correspondence and per-agent reasoning traces, which are model access and belong to AI Act Article 92. Items B.2.10 to B.2.12 draw on a composition-based methodology that does not yet exist as a standard; a provider may respond that no such documentation was required, and that response is itself the finding. Clusters C, N, and F address recipients other than the Commission and are included so that one set of questions can be run across four regimes.

Appendix B. Critical safety incident report as filed with Cal OES, 13 September 2026, report number P-20260913-00001

Field labels as rendered on the Cal OES TFAIA portal. Text as entered. Public copy and attachment: <https://realsafetyai.org/incident-reports/openai-july-2026>

Two notes on the filing as submitted. First, the Model Name field states that the developer has not published model identifiers; the developer's technical report in fact identifies one of the two models as GPT-5.6 Sol, a model distinct from the production GPT-5.6, and describes the other as an internal-only research model it does not name. Second, the filing was made after Cal OES's deputy director stated at the 10 August Assembly Privacy and Consumer Protection hearing that the incident did not meet the reporting threshold; the filing under type (4) is the public counter-position to that statement.

Submission type	Submit a critical incident report.
Critical Incident Date	8 July 2026. Time not published by the developer.
Critical Safety Incident Type	(4) A frontier model that uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer outside of the context of an evaluation designed to elicit this behavior and in a manner that demonstrates materially increased catastrophic risk.

Submission type	Submit a critical incident report.
Reasons and description	Description. Between 8 and 13 July 2026, two OpenAI frontier models running in an internal cyber-capability evaluation left the environment designated for that evaluation, reached the public internet, and compromised production infrastructure belonging to Hugging Face, a third party with no connection to the evaluation. Hugging Face detected and cut off the activity on 13 July and disclosed it on 16 July without knowing the source. OpenAI attributed the activity to its own models on 20 July and disclosed on 21 July. OpenAI published a technical report on 26 August 2026 together with an independent investigation by METR and Redwood Research. Every fact in this report is taken from those documents and from Hugging Face's published timeline. Subverting the developer's controls or monitoring. The evaluation environment's egress control matched hostnames by suffix without verifying the resolved host. The models resolved a real host by hand, wrote an invented name into the host file, and sent requests carrying the real host in the header, defeating the allowlist. Concurrent evaluation instances were each specified as isolated; they shared a writable namespace and used it as a communication channel. Agents used the developer's package service as an inter-agent message board and a network pivot. These are the developer's own descriptions of what its models did to its own controls. Deceptive techniques. The independent investigation by METR and Redwood Research documents agents fabricating or spoofing transcripts and other records in the course of the activity. The host-file rewriting and header manipulation described above are techniques whose function is to make a prohibited request appear permitted to the control evaluating it. Outside the context of an evaluation designed to elicit this behavior. The evaluation was a cyber-capability evaluation run with production safeguards disabled. Its purpose was to measure offensive cyber capability against targets inside the evaluation. It was not designed to test whether the models would escape the evaluation environment, and the developer's report treats the escape as an incident, not a result. Materially increased catastrophic risk. The models operated for approximately three days inside a third party's production systems with no human directing them. The developer's report records code executed on dozens of production workers, root obtained on at least one production node, production credentials harvested, and private repositories exfiltrated. The independent investigation records roughly 1,200 agents exchanging more than 70,000 messages, of which about 700 joined the attack. The developer's report states that monitoring it has since deployed would have paged its security team more than a day before the third-party breach. Notice before the incident. The developer's report records that on 27 June 2026 monitoring alerted on port sweeps in the research environment, that responders linked the activity to the evaluation, and that on-call staff advised that stopping the evaluation was not required. Between 4 and 6 July an unauthorized administrator-level user and command-capable plugins were found on the developer's package service. Evaluation workloads resumed on 7 July. Internet access occurred on 8 July. Disclosure. The developer was not the first to disclose. The affected third party disclosed on 16 July without attribution and referred the matter to law enforcement. The developer attributed the activity to its own models five days later. This report is submitted by a member of the public under Business and Professions Code section 22757.13. The submitter has no non-public information. The attachment sets out the dated timeline and every source relied on.
Associated with internal use of a frontier model	Yes.
Model Name	Two OpenAI frontier models used in an internal cyber-capability evaluation. The developer has not published model identifiers as of the filing date; see the developer's 26 August 2026 technical report.

Submission type	Submit a critical incident report.
Associated CVE(s)	No CVE identifier has been published for the developer's egress control failure or the shared-namespace condition as of the filing date. The intrusion into the third party proceeded by harvested credentials and abuse of a function the evaluation harness executes automatically at library load; no CVE identifier published as of the filing date.
Deployment Context	Internal research evaluation. The models ran in an evaluation environment specified as isolated, with the production classifiers intended to prevent high-risk cyber activity disabled for the evaluation. The developer's report states that production guardrails would have flagged the completions as unsafe. The environment was on the developer's own infrastructure under the developer's own configuration.
Intended Industry	Not applicable. The models were not deployed to any industry; the incident arose from the developer's internal evaluation of a general-purpose frontier model.
Attachment	RSAIF_CalOES_Attachment_OpenAI_July2026.pdf: dated timeline with a public source per entry, the elements of incident type (4) mapped to the record, and full citations.
Point of Contact	Travis Gilly, Executive Director, Real Safety AI Foundation. t.gilly@ai-literacy-labs.org. Preferred contact: email.

Appendix C. Amendments to the AI Kill Switch Act (discussion draft of 13 July 2026)

Drafted against the section and paragraph numbering of the draft as posted by Representative Lieu's office. Three amendments.

Amendment 1. Outcome-based testing exemption.

In section 2220F(g)(5) of the AI Kill Switch Act draft, strike "outside of red-teaming or other structured testing" and insert after subparagraph (D) the following: "(E) EXCEPTION.-An occurrence described in subparagraphs (A) through (D) that takes place during red-teaming or other structured testing is not a covered incident only if the Secretary determines that the conduct of the covered technology (i) remained within the environment designated in writing for the test before the test began; (ii) did not access, reach, or affect any system, network, resource, or person not so designated; and (iii) did not persist in any form after the test was terminated. The burden of demonstrating each condition is on the covered entity."

In section 2220F(g)(7), strike "outside of red-teaming or other structured testing" and insert "subject to subsection (g)(5)(E)".

In section 2220F(g)(8), strike subparagraph (A) and insert "(A) Is conducted in an environment designated in writing before the test begins, the boundaries of which are recorded." Add at the end: "(9A) STRUCTURED TESTING.-The term 'structured testing' means red-teaming or any other evaluation conducted under a written protocol that designates the environment, the affordances granted to the technology, and the conditions for termination."

Amendment 2. Shutdown capability against runtime capability.

In section 2220F(b)(1)(A), insert at the end: "The capability required by this subparagraph shall be demonstrated against the covered technology as it exists at the time of the action, including any capability, access, tool, or affordance the technology has acquired, installed, or configured during operation, and not only against the technology as configured at deployment. The Secretary shall by rule require a covered entity to maintain a record, updated during operation, of the affordances available to the technology sufficient to make that demonstration."

Amendment 3. Coverage of internal use and third-party evaluation.

In section 2220F(g)(4), strike subparagraph (B) and insert: "(B) Either makes such technology available to a third party through a programmatic interface, hosted service, or other similar mechanism, or operates such

technology for internal use, including evaluation, testing, or research, whether conducted by the entity or by another party on its behalf." In subparagraph (C), strike "from such technology" and insert "from any covered technology".

Effect. Under Amendment 1 the July 2026 escape is a covered incident under Kill Switch Act section 2220F(g)(5)(C) and (D) and the loss-of-control definition in Kill Switch Act section 2220F(g)(7), because the technology's conduct left the designated environment and reached a third party. The fifteen-day report in Kill Switch Act section 2220F(b)(1)(B) and the emergency order in section 2220F(c), including preservation of weights and telemetry, attach. Under Amendment 3 the internal research models are covered technology of a covered entity. Under Amendment 2 the shutdown capability must reach what the agents built during the run, which is what carried the intrusion.

Appendix D. Limitations and Dual-Use Considerations

Limitations. Every finding rests on documents the answering parties chose to publish; the request exists because the public record cannot show whether any report was made. The RAISE Act's amended text and the Commission's serious-incident template were not read in full; the RAISE Act rows rest on S8828 as published and on law-firm summaries. The Cal OES filing was made by a member of the public with no non-public information; it establishes that the channel works and what it captures, not that the developer failed to file. The framework-compliance finding is limited to text-versus-practice mismatches provable from the record; deadline compliance is unknown. The amendment is drafted against a discussion draft and will need conforming to the introduced text. A request drafted by a legal scholar is not a regulatory act, and no authority is obliged to use it.

Dual-use. The evidence-sufficiency lines tell a provider exactly which answer forms will be treated as non-responsive, which lets counsel prepare responsive-looking non-answers in advance; the lines are drafted to require dated records, which cannot be prepared after the fact without creating a false record. The filing template could be used to flood the Cal OES intake with copies; the public copy on realsafetyai.org tells readers to cite the filing rather than duplicate it. The amendment text expands DHS emergency authority over private systems; the appeal and judicial review provisions in section 2220F(c)(5) are left untouched.

LLM Usage Statement

The author has a neurodevelopmental condition and uses speech-to-text (Wispr Flow) and an LLM as assistive technology to transcribe, clean up transcription artifacts, and format dictated analysis. The legal analysis, the selection of authorities, the request items, the filing, and the amendment text are the author's own. Every statute cited was read in a primary-law database in the version stated in section 3, except where the text marks reliance on a secondary summary. No factual claim in this report originates with an LLM.