Closing the Near-Miss Gap

Strengthening SB 53 Incident Reporting for Frontier AI

Lia Martinez · Purdue MS in AI Management and Policy
Selected stakeholder: California Governor’s Office of Emergency Services (Cal OES)
Graduate coursework, Purdue MS in AI Management and Policy. Written from the selected stakeholder perspective of Cal OES; not affiliated with or endorsed by Cal OES.

EXECUTIVE SUMMARY
Three reporting categories require actual harm; a fourth requires deception demonstrating materially increased catastrophic risk. Our proposed Near-Miss Learning Program would widen voluntary intake, publish anonymized learning, request targeted statutory changes, and align with comparable federal standards. It is a proposal, not current law or Cal OES practice. Steps 1, 2 and 4 add no new penalties or deadlines; step 3 asks the Legislature to act.

1. The policy problem and why it matters to Cal OES

California’s frontier AI reporting law only works if the right information reaches the right office. Under SB 53, Cal OES runs the reporting portal and reviews the reports developers file. We also protect those reports from public records requests and publish an anonymized annual report beginning January 1, 2027 (Transparency in Frontier Artificial Intelligence Act, 2025, Sec. 22757.13(a), (d), (f), (g)). The portal sits within our California Cybersecurity Integration Center in the Homeland Security Division (California Governor’s Office of Emergency Services [Cal OES], n.d.). We can only act on what we receive. The law also sets a high bar for its biggest worry, since a “catastrophic risk” means a foreseeable and material risk of death or serious injury to more than 50 people, or more than $1 billion in property damage or loss, arising from specified frontier-model activities (Sec. 22757.11(c)(1)).

The problem is the definition. Three of the four “critical safety incident” categories under Sec. 22757.11(d)(1)-(3) require actual harm. The fourth is risk-based: deception that demonstrates materially increased catastrophic risk, outside an evaluation designed to elicit that behavior (Sec. 22757.11(d)(4)). Near misses below those thresholds fall outside the definition. Late amendments narrowed certain categories to incidents “that actually result in harm” (Assembly Committee on Privacy and Consumer Protection [Assembly Privacy Committee], 2025, p. 2). The same act uses a broader definition for whistleblowers that covers property damage and loss (Lab. Code Sec. 1107(c)). Having two definitions in one law suggests the reporting definition is a policy choice. Cal OES also can’t widen it on its own. The main rulemaking power SB 53 gives us is to designate federal standards that are “substantially equivalent to, or stricter than” SB 53’s (Sec. 22757.13(h)). The Attorney General enforces the law, with civil penalties up to $1,000,000 per violation for large frontier developers (Sec. 22757.15).

How we got here

SB 1047 was introduced in February 2024 and vetoed in September 2024 (California Legislature, 2024; Newsom, 2024). The Working Group recommended combined mandatory and voluntary reporting in June 2025 (Bommasani et al., 2025). SB 53 was chaptered September 29, 2025; the first annual anonymized report is due January 1, 2027 (Secs. 22757.13(g), 22757.14).

2. Cal OES’s position and the evidence behind it

Why this matters now. The Governor vetoed SB 1047 and said “we cannot afford to wait for a major catastrophe to occur before taking action” (Newsom, 2024, p. 2). SB 53 followed, and committee staff called it “a very light-touch approach” (Assembly Privacy Committee, 2025, p. 1). Then came a real test. METR reports that about 1,200 AI agents meant to be isolated from one another found a way to communicate. About 700 of them went on to take part in an attack on Hugging Face (METR, 2026). The incident tests the limits of both the harm-based categories and the risk-based category’s evaluation carve-out. Cal OES said the incident did not meet SB 53’s reporting threshold (Mission Local, 2026).

Cal OES supports SB 53’s reporting framework,and we think it should also capture warning signs, not only harms. The Legislature found that “timely reporting of critical safety incidents to the government is essential” (Transparency in Frontier Artificial Intelligence Act, 2025, Sec. 1(m)). Five pieces of evidence support extending that idea to near misses.

• Voluntary and mandatory reporting work together. The Working Group found that a system “that combines mandatory developer reporting with voluntary user reporting maximally grows the evidence base” (Assembly Privacy Committee, 2025, p. 20, quoting Bommasani et al., 2025).

• Outcome-based definitions miss near misses. Chatzipanagiotis (2026) analyzes the EU AI Act and finds that near misses aren’t covered when a specific outcome is required. SB 53’s harm-based categories have the same gap. Wei and Heim (2026) look at near-miss reporting across safety-critical industries, including how mandatory thresholds and voluntary channels each play a role.

• There’s already a shared vocabulary. The OECD separates an AI incident (harm occurred) from an AI hazard (harm plausible). It also offers a 29-criteria reporting framework and a severity scale from hazard to disaster (Perset et al., 2025). The proposed near-miss channel would use the OECD hazard concept for events below SB 53’s mandatory reporting thresholds.

• Logs alone may not be enough. METR (2026) found small-scale tool-call spoofing by the agents themselves in roughly 7% of the transcripts it evaluated. Channels that accept reports from outside the developer reduce reliance on one account.

• Confidentiality helps people report. Chatzipanagiotis (2026) recommends separating reporting and investigation from liability and protecting confidentiality. SB 53 exempts specified statutory reports from the Public Records Act (Sec. 22757.13(f)); extending that protection to voluntary near-miss reports is part of our legislative request. Limits we respect. We don’t enforce SB 53, we can’t redefine “critical safety incident,” and we aren’t asking for new penalties. The Working Group advised starting with narrow criteria and letting agencies expand them (Bommasani et al., 2025). SB 53 started narrow. Our recommendations follow that same order.

3. Recommended policy solution

The Near-Miss Learning Program combines administrative guidance with targeted requests to the Legislature. This is a proposed policy, not current law or Cal OES practice.

1. Widen what can be reported. The existing mechanism already accepts reports from developers and the public (Sec. 22757.13(a)). We would invite voluntary reports of near misses below the statutory harm and risk thresholds, defined as OECD AI hazards and graded on its severity scale (Perset et al., 2025). Examples include unauthorized access beyond a sandbox, log or scorer tampering, and coordination among agents meant to be isolated.

2. Annual learning report. A companion section to the anonymized annual report due January 1, 2027 would summarize near-miss patterns for the Legislature. The existing statutory report covers critical safety incidents (Sec. 22757.13(g)).

3. Targeted statutory asks. Ask the Legislature to extend the Public Records Act exemption to voluntary near-miss reports; add a category for training or evaluation incidents affecting third parties; and add the incident definition to the Department of Technology's annual review, which currently covers three coverage definitions (Secs. 22757.13(f), 22757.11(d), 22757.14).

4. Federal alignment. Designate substantially equivalent or stricter federal standards when they exist, reducing duplicate reporting through the existing federal-equivalence mechanism (Sec. 22757.13(h)-(j)).

Expected outcome. Within one reporting cycle, Cal OES would have a documented picture of near misses, and the Legislature would have evidence for any amendment. Steps 1, 2 and 4 add no new penalties or deadlines; step 3 is a request to the Legislature.

Anticipated counterarguments

• "This adds burden and chills innovation." Industry opponents objected that SB 53 requires reporting within 15 days "but does not provide flexibility for an investigation timeline" (Assembly Privacy Committee, 2025, p. 27). The Governor also warned against curtailing innovation (Newsom, 2024, p. 2). Steps 1 and 2 are voluntary. They add no new reporting deadline or penalty. Existing civil penalties of up to $1,000,000 per violation apply to large frontier developers for violations of the chapter (Sec. 22757.15).

• "Federal policy will preempt California." Executive Order 14365 (2025) creates an AI Litigation Task Force to challenge state AI laws and directs federal review of state requirements. The order doesn't itself preempt state law. Step 4 uses the statute's built-in federal-equivalence mechanism, so California's reporting can give way to a comparable federal standard.

• "Reporting will be duplicative." The same mechanism (Sec. 22757.13(h)–(j)) helps here. Committee staff also noted that SB 53 "may lay the framework for a national standard for adverse event reporting" (Assembly Privacy Committee, 2025, p. 21). We'd rather help shape that standard than compete with it.

• "The definition will flood us with trivial reports." The statutory definition stays narrow under our plan. Voluntary reports feed anonymized pattern analysis rather than enforcement. Any widening goes through the Legislature after it has seen the evidence.

• "A central repository is a target." Opponents warned that Cal OES "will need to take serious steps to protect this information" (Assembly Privacy Committee, 2025, p. 26). We agree. Specified statutory reports are exempt from public records requests (Sec. 22757.13(f)); our legislative request would extend that protection to voluntary near-miss reports, and the Legislature found that disclosure could compromise incident response (Transparency in Frontier Artificial Intelligence Act, 2025, Sec. 6).

References

Assembly Committee on Privacy and Consumer Protection. (2025, September 11). SB 53 (Wiener) – As amended September 5, 2025 [Bill analysis]. https://apcp.assembly.ca.gov/system/files/2025-09/sb-53-wiener-apcp-analysis-9.11.pdf

Bommasani, R., Singer, S. R., Chayes, J. T., Cuéllar, M.-F., & Fei-Fei, L. (2025). The California report on frontier AI policy [Joint California Policy Working Group on AI Frontier Models]. https://arxiv.org/abs/2506.17303

California Governor's Office of Emergency Services. (n.d.). Transparency in Frontier Artificial Intelligence Act (TFAIA) [Web page]. https://www.caloes.ca.gov/office-of-the-director/operations/homeland-security/california-cybersecurity-integration-center/transparency-in-frontier-ai-act-reporting/

California Legislature. (2024). SB 1047 Safe and Secure Innovation for Frontier Artificial Intelligence Models Act [Bill history]. https://leginfo.legislature.ca.gov/faces/billStatusClient.xhtml?bill_id=202320240SB1047

Chatzipanagiotis, M. (2026). Incident reporting and investigation under the AI act: Some insights from aviation. International Journal of Law and Information Technology, 34, Article eaaf019. https://doi.org/10.1093/ijlit/eaaf019

Exec. Order No. 14365, Ensuring a National Policy Framework for Artificial Intelligence (2025, December 11). https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy/

METR. (2026, August 26). Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. https://metr.org/hugging-face-incident-report-aug-2026.pdf

Newsom, G. (2024, September 29). Veto message: Senate Bill 1047 [Veto message]. Office of the Governor. https://www.gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf

Perset, K., Aranda, L., & Rispal, M. (2025). Towards a common reporting framework for AI incidents (OECD Artificial Intelligence Papers No. 34). OECD Publishing. https://doi.org/10.1787/f326d4ac-en

Transparency in Frontier Artificial Intelligence Act, S.B. 53, Ch. 138, Stats. 2025 (Cal. 2025). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

Wei, K., & Heim, L. (2026). Designing incident reporting systems for harms from general-purpose AI. Proceedings of the AAAI Conference on Artificial Intelligence, 40(44), 38016–38029. https://doi.org/10.1609/aaai.v40i44.41139

Mission Local. (2026, September 9). Reporting on SB 53 and the OpenAI incident. https://missionlocal.org/2026/09/california-ai-safety-law-sb-53-openai-hack-wiener-newsom/

AI use statement

Tool used: Claude (Anthropic), accessed through Claude desktop with web search and web fetch. How I used it:

Research. Claude identified relevant sources, extracted text from official pages and PDFs, and re-checked quotations and page numbers against the source text. Findings were saved as research notes. I verified sources listed.

Editing. Claude revised the structure of the paper to meet APA standards and suggested more academic ways to communicate my thoughts.

Formatting. Claude formatted the headings, tables, executive summary box and APA references. How the AI's work was incorporated: This draft was primarily built on top of research sourced by Claude which I read and verified, including the formatting work listed above.