AMERICAN SOCIETY FOR AI

Verifying Frontier AI

Considerations for California's Expert Group · Submitted to the California Government Operations Agency under Executive Order N-9-26 · November 5, 2026

Executive Summary

The **American Society for AI** (ASFAI) is a private, non-partisan 501(c)(3) non-profit founded in 2023 that advocates for the safe and beneficial advancement of AI led by the United States. Its members include Fortune 100 executives, members of Congress, and preeminent AI researchers and entrepreneurs.

California's **Executive Order N-9-26** directs the Government Operations Agency (GovOps), along with the Governor's Office of Emergency Services, to develop recommendations with national experts. This submission discusses advantages and disadvantages of several proposals. The proposals are for discussion only and do not represent the official position of ASFAI.

- **Question 1. Embedded verification:**
  - Proposal 1.1: Verifiers should have continuous mandatory access.
  - Proposal 1.2: Custody of the verification records should be shared.
  - Proposal 1.3: Verifiers should be independent of agencies that promote AI.
- **Question 2. Verified frameworks:**
  - Proposal 2.1: Verification should extend to deployed systems.
  - Proposal 2.2: Each model should have deployment guidelines.
  - Proposal 2.3: Deployments within a model's guidelines should get a safe harbor.
- **Question 3. Kill switch:**
  - Proposal 3.1: A kill switch should be a containment switch.
  - Proposal 3.2: AI chips should not include a stop mechanism.
  - Proposal 3.3: A kill switch requirement should depend on deployment risk.
- **Question 4. Critical safety incidents:**
  - Proposal 4.1: Incidents should include near misses and training and evaluation incidents.
  - Proposal 4.2: Incident investigations should be blame-free.
  - Proposal 4.3: Agent actions should be logged with agent and responsible human identity.
- **Definitions**
- **Historical Evidence**

Key contributors: Michael Carey, Amyn Jan, Cyril Gorlla, Lena Smart, Russ Wilcox.

Declaration of interest: Cyril Gorlla is Chief Executive Officer of CTGT, which published the pipeline study cited in Section 2.

1. Embedded Verification Organizations

Question

The first question asks the State to "Require frontier AI companies to embed a designated independent verification organization onsite in their labs to conduct regular audits and evaluations."

Record

The most prominent independent investigation of a frontier AI incident so far, into the Hugging Face attack (see Historical Evidence), took place on site but on the developer's terms.[1] METR staff and a Redwood Research contractor spent six days on OpenAI's premises under a scope agreed with OpenAI that left out the effectiveness of safeguards. OpenAI set the investigation window, which closed on July 13 although message-board entries run to July 19, and investigators could not query the model most involved, which OpenAI says it deactivated and encrypted. METR told a Senate subcommittee that "the participation of these AI developers is voluntary".[2]

Other regulators keep the safety judge apart from the promoter. In 1974 the US moved nuclear safety regulation from the Atomic Energy Commission, which also promoted nuclear power, to the new Nuclear Regulatory Commission.[3] China's National Nuclear Safety Administration is likewise separate from the administration that promotes nuclear power.[4] China's grid is an exception: the energy administration that promotes AI pilots in the power sector also supervises grid security.[5] (See Historical Evidence.)

Framework

Location. An auditor on the developer's premises sees systems before release and can move faster than one outside. Presence alone does not settle scope, as the Hugging Face case shows.

Record owner. A verifier checks the records it holds (see Call-level record under Definitions). If one party holds the only copy, that party can set the terms of access after something goes wrong.

Licensing. The verifier could be a public body, a private verifier licensed by government,[6] or a self-regulatory body that sets standards, as FINRA does for broker-dealers.

Independence. A public verifier would be paid by the state and should not sit within an agency that regulates or promotes AI; a private verifier would be paid by the developers. Either way, who pays raises a question of independence.

Proposals

Proposal 1.1: Verifiers should have continuous mandatory access. Verifiers would have standing access that the developer cannot narrow, including staff resident at each frontier developer, as the order proposes.

> **For.** Only an onsite verifier sees systems before release, and voluntary access can be narrowed at will, as OpenAI did with the window and scope of the Hugging Face investigation. After

[1] METR, independent investigation of the OpenAI and Hugging Face incident, blog post and full report, Aug 26, 2026. metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation; metr.org/hugging-face-incident-report-aug-2026.pdf
[2] Chris Painter (METR), testimony to the Senate HSGAC Subcommittee on Disaster Management, District of Columbia, and Census, Sept 30, 2026. metr.org/blog/2026-09-30-chris-painter-senate-testimony
[3] NRC history, Energy Reorganization Act of 1974. nrc.gov/about-nrc/history.html
[4] National Nuclear Safety Administration, statement on two operating events at Haiyang NPP Unit 2, Mar 22, 2020. www.mee.gov.cn/ywdt/gsgg/tz/202003/t20200322_769961.shtml
[5] NDRC Order 27, power monitoring system security rule (adopted Nov 25, 2024; in force Jan 1, 2025). ndrc.gov.cn/xxgk/zcfb/fzggwl/202412/P020241211585903621428.pdf; National Energy Administration, Notice on the List of High-Value "AI+" Energy Scenarios and Pilot Applications (国能综通科技〔2026〕50号), May 14, 2026. www.nea.gov.cn/20260526/261bbb10eea24be3b7ea3a00d89b8891/c.html
[6] Gillian Hadfield and Jack Clark, Regulatory Markets: The Future of AI Governance, 2023. arxiv.org/pdf/2304.04914

Three Mile Island the NRC assigned resident inspectors to each nuclear plant, at least two of whom live nearby and work only at that plant.[7]

Against. A resident auditor risks capture by the developer it checks; onsite access to weights adds security exposure; and the cost falls hardest on smaller developers.

Open questions. How resident-inspector and licensed-examiner arrangements in other regulated sectors set access, scope, rotation and record custody, and what they have caught.

Proposal 1.2: Custody of the verification records should be shared. Developers, deployers, the communities where agents are deployed, and government would each hold a copy of the call-level record. One version would pair an automated monitor of exchanges between agents with community councils that hear complaints and publish findings without a veto.

For. No single party would hold the evidence. The International Labour Organization seats governments, employers and workers side by side, and China's grid shows that a second copy matters only when its holder is independent of the party being checked.

Against. Each added copy of call-level records adds privacy, trade-secret and security exposure and storage cost, and split custody does not reach an open-weight model that runs with no developer in the loop.

Proposal 1.3: Verifiers should be independent of agencies that promote AI. Verification would sit outside both the developer and any agency that promotes AI adoption, in a public body or in private verifiers that government licenses.

For. Whoever promotes adoption should not be the sole judge of safety, the premise of the NRC's 1974 separation from the Atomic Energy Commission and of China's separate nuclear regulator. Licensing private verifiers lets the government set the standard while the market supplies the capacity.[8]

Against. A verifier with no stake in deployment can drift toward caution at no cost to itself: in May 2025 an executive order told the NRC to add facilitating nuclear power to its mission and to decide new reactor applications within 18 months, faulting "a myopic policy of minimizing even trivial risks".[9] A licensed private verifier paid by developers raises its own question of independence.

[7] US Nuclear Regulatory Commission, "Backgrounder on the Three Mile Island Accident." www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle
[8] Gillian Hadfield and Jack Clark, Regulatory Markets: The Future of AI Governance, 2023. arxiv.org/pdf/2304.04914
[9] Executive Order 14300, Ordering the Reform of the Nuclear Regulatory Commission, May 23, 2025. govinfo.gov/content/pkg/DCPD-202500633/pdf/DCPD-202500633.pdf

2. Verified Frameworks and Reports

Question

The second question asks the State to "Require that the AI safety frameworks, transparency reports, and risk assessments" be "verified pursuant to standards deemed adequate by an independent verification organization."

Record

SB 53 requires large frontier developers to publish safety frameworks and transparency reports that describe "the role of any third-party evaluators," and it dropped SB 1047's annual third-party audit.[10] California's 2025 frontier-policy report called third-party assessment "essential" and proposed safe harbors for evaluators "analogous to those afforded to third-party cybersecurity testers".[11]

The White House Accord on Super Intelligence (September 29) asks developers to have an independent evaluator assess whether their controls are "operating as intended",[12] and leaves auditor selection, independence criteria and access to the developers.[13] Neither the Accord nor the order defines intended operation in terms a third party can test, and neither says who holds the record such a check depends on.

One published study shows why a check of the model does not answer for the system around it. A model returned identical text for identical input, yet the same policy manual run through the surrounding pipeline four times produced four different sets of obligations, and one run never wrote an escalation rule. One faulty request gave the same wrong answer 8 of 8 times, and a repeat-and-compare check passed it. The dropped rule was traced to its cause only because every model call and prompt had been recorded.[14]

Aviation has required such a record for decades: the flight recorder is built to survive the crash, and after an accident an investigator independent of the airline takes custody of it (see Historical Evidence).

Framework

Systems. Verifying documents checks what a developer says about a system; re-running its tests checks the system. Alignment safeguards and deployment safeguards are controlled by different parties and fixed at different times, so a check of one does not answer for the other. A model can be deterministic while the deployed system is not, and a repeat-and-compare check catches variance but misses an error that repeats, so it needs to be paired with checks against a statement of intent.

Proportionality. The US, Canadian and UK nuclear regulators' joint principles paper treats written documentation as central to showing an AI system is adequately safe, and frames management around the consequences of failure and the level of autonomy.[15] On that view a customer chatbot and a credit-decision agent would not be treated alike.

Requirements. Members identified five elements:

---

[10] Justine Gluck, "California's SB 53: The First Frontier AI Law, Explained," Future of Privacy Forum, Oct 3, 2025 (updated Feb 4, 2026). fpf.org/blog/californias-sb-53-the-first-frontier-ai-law-explained/
[11] Joint California Policy Working Group on AI Frontier Models, The California Report on Frontier AI Policy, June 17, 2025. arxiv.org/html/2506.17303v1
[12] White House Accord on Super Intelligence, Sept 29, 2026, item 3, text via The American Presidency Project. www.presidency.ucsb.edu/documents/white-house-accord-super-intelligence
[13] Jennifer Everett, Cynthia Cole and Santi Villar, "White House Secures Voluntary Industry Commitments on Independent AI Audits," Alston & Bird Privacy, Cyber & Data Strategy Blog, Oct 1, 2026. www.alstonprivacy.com/white-house-secures-voluntary-industry-commitments-on-independent-ai-audits/
[14] CTGT, "Deterministic Inference Did Not Make Our Pipeline Deterministic" (Concordance), Oct 21, 2026. ctgt.ai/research/concordance
[15] US NRC, CNSC and UK ONR, Considerations for Developing Artificial Intelligence Systems in Nuclear Applications, September 2024. nrc.gov/docs/ML2424/ML24241A252.pdf

- A statement of intent for both the model and its deployment, since an outsider can test only what was written down before the fact.
- A test an outsider can re-run on the same build and reach the same verdict.
- Re-testing whenever the system changes, including a new model behind the same product name or a new deployment configuration.
- A recorded verdict when a check was not performed, so a missed check never reads as a pass. Nuclear technical specifications treat a missed surveillance as failure to meet the operating limit, subject to a narrow exception.[16]
- Call-level records the operator cannot alter.

Intent. Agents act in ways no one can list in advance, so intended behavior cannot be written as a list of permitted actions. Nuclear technical specifications offer a pattern: limiting conditions a system must stay within and how often it must be shown to be within them. A written statement of intent would describe that envelope: what must never happen, what must always hold, and what the system may touch.

Agent terms. Members also raised writing the envelope as a social contract for agents: standing terms binding people, agents, developers and deployers, instead of a constitution each developer drafts for its own model. The terms would set which decisions only the person may make, which need the person's authorization and which the agent may make alone; give the person an account of what the agent did that the person may contest; let the agent decline a request that would harm someone else; treat prompt injection that hijacks an agent as seizing the will of the person it acts for; and keep whoever profits from an agent from controlling what it presents as fact. Supporters say such terms give a verifier a testable statement of intent, written down before the fact. Critics answer that standing to refuse could shift liability away from developers. China's May 2026 agent policy writes similar decision tiers into national rules,[17] which shows a second risk: terms that bind agents can also bind the people they serve when one authority holds every record.

Disclosure. Disclosure duties could apply to all developers rather than only those above a compute threshold: each would publish its frameworks, evaluation results and incidents, extending what SB 53 already requires of large frontier developers (see Record). Securities law has worked this way since 1933: the SEC requires that investors receive key information but does not judge the merits of offerings.[18] What a developer discloses is what a verifier can check (Proposal 2.1), and disclosure already supplies part of the record: METR documented 44 incidents by May 2026, and OpenAI disclosed in September that its agents used Census Data API keys found in public repositories (Section 4). Supporters say universal disclosure gives verifiers the data any later threshold would need and is harder for incumbents to bend; some members see calls for strict regulation from large developers as a bid for regulatory capture. Critics answer that disclosure alone does not stop a harmful system from shipping, while a compute threshold is cheap to check and enforce. A middle course would apply disclosure to all developers and keep a compute threshold only as a first filter for added duties, since model size is a poor proxy for risk (Proposal 3.3). Who may redact a disclosure is open: OpenAI's redactions of the METR report were limited, but they were OpenAI's to make (see Historical Evidence).

Proposals

Proposal 2.1: Verification should extend to deployed systems. Verifiers would re-run the tests a framework describes on the same build and check the deployed system against its statement of intent, using call-level records held outside the operator's control.

[16] Technical Specifications, H.B. Robinson Steam Electric Plant Unit 2, Section 3.0, SR 3.0.1 (NRC ADAMS ML050530105). www.nrc.gov/docs/ML0505/ML050530105.pdf
[17] Cyberspace Administration of China, NDRC and MIIT, Implementation Opinions on the Standardized Application and Innovative Development of AI Agents, May 8, 2026. www.cac.gov.cn/2026-05/08/c_1779979789523320.htm
[18] US Securities and Exchange Commission, "Registration Under the Securities Act of 1933," Investor.gov glossary. www.investor.gov/introduction-investing/investing-basics/glossary/registration-under-securities-act-1933

For. Only a check of the running system shows whether it behaves as described. Without call-level records no one can tell a model failure from a deployment failure, and an investigation regime has nothing reliable to learn from.

Against. Document verification is feasible now and scales; behavioral access carries trade-secret and security costs; call-level logging of prompts and outputs raises privacy concerns and adds storage cost; and periodic testing of the model against published evaluations is enough for most deployments.

Open questions. Whether a verifier can re-run the tests a framework describes, on the same build, and reproduce the reported results; whether a verification made on one date says anything about a system that keeps changing; and whether two investigation teams given the same incident, one with call-level records and one with model evaluations only, reach the same cause and the same layer.

Proposal 2.2: Each model should have deployment guidelines. A technically capable agency, such as NIST's Center for AI Standards and Innovation, would test frontier models and issue model-specific deployment guidelines, similar to the operating limitations in an aircraft's type certificate: the autonomy, tool and network access, monitoring and isolation appropriate for a given model. Ways to keep testing from becoming a bottleneck include:

- shipping untested models under the most restrictive default guidelines until testing is done;
- certifying the model family and the developer's process, with audited developers testing their own updates under agency oversight, as holders of an FAA Organization Designation Authorization issue certificates and conduct inspections on the FAA's behalf;[19]
- triggering new testing only for updates that materially change dangerous capabilities against defined evaluations;
- fixed review deadlines, with proposed guidelines taking effect if the agency misses them;
- carrying findings from investigations and near-miss reports to every deployer of the affected model, as FAA airworthiness directives do.

For. Scrutiny scales with the model's capabilities and the deployment's autonomy, and deployers know in advance where they stand.

Against. Developers update deployed systems every few weeks, and agency testing before each change could slow development or turn into approval without review.

Proposal 2.3: Deployments within a model's guidelines should get a safe harbor. Deployers operating within the guidelines would be protected from liability for harms the guidelines did not anticipate; those exceeding them would carry the risk.

For. A safe harbor gives deployers a reason to follow the guidelines and a clear line between compliant and noncompliant use. California's 2025 frontier-policy report proposed a comparable safe harbor for evaluators (see Record).

Against. A safe harbor could shield a deployer that meets the letter of the guidelines while causing foreseeable harm, and it moves the remaining risk onto the people harmed.

[19] Federal Aviation Administration, Organization Designation Authorization (ODA) Office. www.faa.gov/about/office_org/headquarters_offices/avs/offices/oda; 14 CFR part 183, subpart D, Organization Designation Authorization. www.law.cornell.edu/cfr/text/14/part-183/subpart-D

3. Kill Switch

Question

The third question asks the State to "Advance the creation of a 'kill switch' for frontier models," with its efficacy "verified on an ongoing basis by an independent verification organization."

Record

In the Hugging Face case no stop was used. Trajectory monitoring was not in place, isolation relied on a single layer of network controls, and OpenAI did not connect the agents' activity to Hugging Face until about July 20, a week after the attack ended.[20]

California's legislature passed a shutdown requirement once. SB 1047 would have required developers to be able to promptly enact a full shutdown; the Governor vetoed it in September 2024, writing that it did not account for whether a system is deployed in high-risk settings, makes critical decisions or uses sensitive data.[21] Its successor, SB 53, carries no shutdown requirement and instead requires loss-of-control incidents to be reported within 15 days, or within 24 hours when lives are at imminent risk (Section 4).

At the chip layer, neither government requires a stop. Washington's Chip Security Act, passed 42 to 0 by the House Foreign Affairs Committee in March 2026 and not yet law, would require location verification on exported AI chips, and an amended draft states that it may not be read to require a kill switch;[22] Nvidia's security chief has written that its GPUs do not and should not have kill switches.[23] Beijing's cyberspace regulator summoned Nvidia on July 31, 2025 over fears that its China-market chips could be tracked and shut off remotely.[24]

China's standards body has written a stop into its guidance. TC260 guidance, issued in draft in January 2026 and in final form in September 2026, asks operators of AI applications to plan one-click takeover, version rollback and emergency shutdown, and, for systems that interact physically with people, a shutdown a person can easily operate.[25] The operator holds that stop. Beijing is also building a scheduling layer for public compute across chip vendors, with identifiers issued through a national node and monitoring under unified rules,[26] though the national data bureau said in January 2024 that effective scheduling was not yet solved.[27] In the United States, hosted compute rests mostly with three cloud providers, which together held about 63 percent of worldwide spending on cloud infrastructure

[20] Jeremy Kahn and Emily Forlini, "OpenAI says its AI models escaped control and hacked into AI company Hugging Face," Fortune, July 21, 2026. fortune.com/2026/07/21/openai-says-ai-models-escaped-control-hacked-hugging-face; Lorenzo Franceschi-Bicchierai, "How OpenAI's human mistake led to the AI-powered hack on Hugging Face," TechCrunch, July 22, 2026. techcrunch.com/2026/07/22/how-an-openais-human-mistake-led-to-the-ai-powered-hack-on-hugging-face/

[21] Governor's veto message on SB 1047, Sept 29, 2024. gov.ca.gov/wp-content/uploads/2024/09/SB-1047-Veto-Message.pdf

[22] Chip Security Act, H.R. 3447, 119th Cong., www.congress.gov/bill/119th-congress/house-bill/3447, and S. 1705, 119th Cong., www.congress.gov/bill/119th-congress/senate-bill/1705; Office of Rep. Bill Huizenga, "UNANIMOUS: Bipartisan Huizenga Legislation to Curb AI Chip Smuggling Passes Committee," Mar 26, 2026. huizenga.house.gov/news/documentsingle.aspx?DocumentID=404259; draft amendment in the nature of a substitute, July 15, 2025, rule of construction. amendments-rules.house.gov/amendments/Chip%20Security%20Act%20DRAFT%20ANS%207.15.25250828112622351.pdf

[23] David Reber Jr. (Nvidia), "No Backdoors. No Kill Switches. No Spyware," Aug 5, 2025. blogs.nvidia.com/blog/no-backdoors-no-kill-switches-no-spyware/

[24] Cyberspace Administration of China, statement on summoning Nvidia over H20 vulnerability and backdoor security risks (国家互联网信息办公室就H20算力芯片漏洞后门安全风险约谈英伟达公司), July 31, 2025. www.cac.gov.cn/2025-07/31/c_1755675743897163.htm

[25] TC260 Secretariat, Cybersecurity Standard Practice Guide: AI Application Security Guidelines, General Provisions (网络安全标准实践指南——人工智能应用安全指引 总则), TC260-PG-20268A, Sept 15, 2026 (draft for comment Jan 13, 2026), §7.1g and §7.5d. www.tc260.org.cn/tc260/tzgg/202609/ec149159dcb246268eda751eee89967d.shtml

[26] MIIT General Office, Notice on Building National Computing-Power Interconnection Nodes (工信厅信管函〔2026〕35号), Jan 27, 2026. www.miit.gov.cn/jgsj/xgj/wjfb/art/2026/art_67864b944859 43ad8cee03f4f4989d17.html

[27] Xinhua, interview with the principal official of the National Data Administration on building a national integrated computing network, Jan 3, 2024. www.nda.gov.cn/sjj/zwgk/zcjd/0830/20240830193716443740817_pc.html

services in the second quarter of 2026,[28] and the 2025 AI Action Plan encourages open-weight models,[29] which no provider or developer can recall once released.

Outside AI, operators have kept control when they held a fallback or a local cut-off of their own, as Australia's grid operator did in 2025 and Norway's bus operator Ruter did the same year (see Historical Evidence).[30]

Framework

Scope. A kill switch is a narrow tool. It fits frontier models in development or provisional release, but not deployed or open-weight agents, so the scope of a deployment matters as much as its risk.

A kill switch assumes one operator stopping one or more machines under their control; agents act in groups, run on weights no developer controls, and can finish an action before anyone reaches for a stop, so containment describes the task better than a switch. On this view the first question is the layer where containment sits. At the agent layer, written terms would bind agents to one another and to the people they act for (see Agent terms, Section 2), and a monitoring agent would halt a peer once set criteria are crossed in its reasoning; researchers call this trusted monitoring,[31] and proposals for agent IDs, identity binding and rollback supply the technical means.[32]

Containment at the chip layer could reach open-weight models, though a stop written for one vendor's software stack misses chips built on another, and Chinese AI chips run on stacks of their own, including Huawei's CANN[33] and Moore Threads' MUSA, which the company said in March 2024 has no dependency on CUDA.[34] A hard stop fits best where harm cannot be undone; a lab aggregating biological data for wet-lab work should keep its sandbox in place at all times, and an outside party should be able to confirm that it did. In Beijing, BAAI and Peking University found that all 11 commercial models they tested produced plans for splitting DNA orders to get around synthesis screening.[35]

One option is a containment switch (see Definitions) with intermediate levels of response. The switch is independent: it does not depend on the agent recognizing its own failure, agreeing to stop or cooperating. It sits outside the agent's authority: at identity and authorization, tool gateways, the orchestration layer, the execution runtime and the downstream services that carry out the action. Stopping an agent's reasoning loop is not enough if its credentials stay valid, agents it delegated to keep working, or an outside service keeps executing a request it has already accepted. Containment on this view has four tasks: prevent further actions, interrupt actions in progress, isolate connected processes and delegated agents, and limit the consequences of actions already completed, some of which cannot be reversed.

A containment switch sets the point of intervention according to three factors: the authority delegated to the agent, the potential consequences of its actions, and the time available before those consequences become material. It does not rest on a single anomaly or risk score. A numeric cutoff that has not been validated against the operating environment gives false precision, and where harm

[28] Synergy Research Group, "Q2 Cloud Market Passes $143 Billion; Highest Growth Rate in Eight Years," July 30, 2026 (Amazon 28%, Microsoft 20%, Google 15%). www.srgresearch.com/articles/q2-cloud-market-passes-143-billion-highest-growth-rate-in-eight-years
[29] Winning the Race: America's AI Action Plan, July 2025. whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf
[30] Australian Energy Market Operator, "Self-forecasting errors and frequency excursion on 19 August 2025," Reviewable Operating Incident Report, May 11, 2026 (v1.1, Sept 21, 2026). www.aemo.com.au/-/media/files/electricity/nem/market_notices_and_events/power_system_incident_reports/2025/self-forecasting-errors-and-frequency-excursion-on-19-august-2025.pdf
[31] Greenblatt et al., AI Control, 2023. arxiv.org/pdf/2312.06942
[32] Chan et al., Infrastructure for AI Agents, 2025. arxiv.org/pdf/2501.10114
[33] Xinhua, "Huawei open-sources chip software ecosystem," via China Daily, Aug 7, 2025. www.chinadaily.com.cn/a/202508/07/WS68940c32a3108a99c19059fd.html
[34] "英伟达禁止第三方使用CUDA？摩尔线程：旗下产品不涉及，开发者可放心使用" (Moore Threads statement on MUSA), The Paper, Mar 6, 2024. www.thepaper.cn/newsDetail_forward_26574316
[35] Beijing Daily on the BAAI and Peking University study, July 22, 2026. news.bjd.com.cn/2026/07/22/11880629.shtml

would be severe and irreversible, uncertainty alone can justify intervening. The response would be graduated:

Level	When	Response
1. Observe	Anomalous behavior or a departure from the expected pattern, with no credible evidence of a policy violation or material harm	Increase monitoring, preserve records, assess
2. Restrict	The agent approaches or attempts to cross a defined boundary of authority	Restrict tools or permissions, limit execution, require authorization before consequential actions; the agent keeps working within the authority that remains
3. Contain	A policy violation is confirmed, an unauthorized consequential action is attempted, or a credible risk of material harm cannot be controlled by ordinary restrictions	Suspend the workflow, revoke permissions, interrupt tasks where possible, isolate the agent from the systems it could affect
4. Emergency containment	Severe harm is imminent, a significant compromise is confirmed, or containment is failing	Terminate execution, revoke credentials, isolate dependent agents and services, block consequential actions across the environment, without waiting for a human operator if delay would increase the harm
5. Systemic shutdown	The threat extends beyond one agent or workflow, and the integrity of the wider environment cannot be assured	Disable agent execution across the environment, isolate shared infrastructure, start formal incident response and recovery

Level 5 is the closest equivalent to the traditional kill switch. Each level needs activation criteria, an enforcement mechanism, escalation rules and conditions for restoring authority. The rule across levels is to intervene at the lowest level that reliably contains the risk and escalate when it does not. Accountable system owners set the limits, not the agent, and the limits are enforced in controls, not in a system prompt. A parent agent's shutdown does not stop its children or the agents it has interacted with, so containment works at each agent's identity, permissions and execution environment, and fails closed for designated high-risk actions when enforcement cannot confirm they are safe. Because an agent can act faster than monitoring detects, irreversible or high-impact actions, such as a financial transaction or a destructive change to a production system, need authorization at the point of execution. Where the architecture allows, transaction boundaries, cancellation, rollback and compensating actions limit what has already happened downstream.

Reach. A switch reaches only systems the developer still controls. Released open weights cannot be recalled, so efficacy has to be stated per deployment, not per model.

Authority. China defines national security broadly in statute;[36] US law defines it piecemeal, with no definition written for AI, and a December 2025 executive order directs the Justice Department to challenge state AI laws,[37] so it is an open question whether a town, a state or the federal government would hold the authority to stop a system. Beijing's scheduling layer would place a stop in one authority's hands, while in the United States no single party holds a layer that reaches every deployed agent.

[36] National Security Law of the People's Republic of China, Presidential Order No. 29, July 1, 2015, Art. 2. www.gov.cn/zhengce/2015-07/01/content_2893902.htm; English translation: govt.chinadaily.com.cn/s/201812/11/WS5c0f1b56498eefb3fe46e8c9/national-security-law-of-the-peoples-republic-of-china-2015-effective.html

[37] Executive Order 14365, Ensuring a National Policy Framework for Artificial Intelligence, Dec 11, 2025, 90 Fed. Reg. 58499 (Dec 16, 2025). www.federalregister.gov/documents/2025/12/16/2025-23092/ensuring-a-national-policy-framework-for-artificial-intelligence

Proposals

Proposal 3.1: A kill switch should be a containment switch. A kill switch requirement would be met by the five containment levels above, with limits set by accountable system owners and enforced in controls, and with efficacy stated per deployment and measured by time to detection, time to enforcement, the number and severity of consequential actions possible before containment, and the authority the agent retains after intervention.

> **For.** A shutdown capability depends on the system around it: defined authority to halt, triggers set in advance, monitoring that can see the trigger, and automatic halts on anomalies within compute, time and action budgets. The NRC gap assessment[38] asks for detection of inputs unlike the training data and of anomalous outputs, and mitigation so errors do not propagate; carried over to general AI, that means safe states defined before go-live. In Australia's grid case, the operator's own fallback let an automated failure end safely.
>
> **Against.** A built-in shutdown path is a vulnerability that good and bad actors could both use: "It becomes an engineering problem rather than a rogue AI problem".[39] A monitoring agent is itself an attack surface. Agents can pass poisoned instructions from one to the next,[40] and researchers who poisoned a driving agent's memory chose a dangerous sudden stop as one attack goal, succeeding more than 80 percent of the time with under 0.1 percent of the memory poisoned.[41] Triggers read from a model's reasoning work only while that reasoning stays readable, which researchers at the major developers call fragile.[42] A monitor able to rewrite another agent's context also owes an account others can see and contest.
>
> **Open questions.** Whether a stop halts a test system within a set time, including when the system or a peer tries to evade it; how often it is tested, and whether someone other than the operator runs the test; how many of the 44 incidents METR documented a containment trigger set in advance would have caught; and the four measures above, tested with failed communication, unresponsive agents, compromised credentials, concurrent execution and actions already in flight.

Proposal 3.2: AI chips should not include a stop mechanism. AI chips would not carry a mechanism that could halt a model wherever it runs; stops would sit at the operator, deployment and agent layers instead.

> **For.** A stop written for one vendor's software stack misses chips built on another, including Chinese chips on stacks of their own. Neither the Chip Security Act nor Nvidia supports a chip-level stop, and Beijing treats one as a threat (see Record).
>
> **Against.** Containment at the chip layer could reach open-weight models, which no provider or developer can recall once released.
>
> **Open questions.** Whether a chip-level mechanism can be verified by both governments as not a backdoor.

Proposal 3.3: A kill switch requirement should depend on deployment risk. Whether a system must have a kill switch would depend on its deployment risk (autonomy, tool access and the consequences of failure) rather than on the size of the model.

---

[38] Southwest Research Institute for the NRC, Regulatory Framework Gap Assessment for the Use of Artificial Intelligence in Nuclear Applications, October 2024. nrc.gov/docs/ML2429/ML24290A059.pdf
[39] Madeline Shannon, "California governor convenes AI panel to ponder 'kill switch'," The Center Square, Oct 7, 2026. www.thecentersquare.com/california/article_35f31e87-3509-4491-8628-182bb8b6dd11.html
[40] Lee and Tiwari, Prompt Infection, 2024. arxiv.org/abs/2410.07283
[41] Chen et al., AgentPoison, 2024. arxiv.org/pdf/2407.12784
[42] Korbak et al., Chain of Thought Monitorability, 2025. arxiv.org/pdf/2507.11473

For. The Governor's 2024 veto message made a related point, that smaller, specialized models could prove as dangerous as large ones, and Shanghai AI Laboratory's framework judges risk by deployment environment, threat source and enabling capability against yellow and red lines, not by compute.[43]

Against. A compute threshold is cheap to check and makes obligations enforceable.

[43] Shanghai AI Laboratory, Frontier AI Risk Management Framework, 2025. arxiv.org/pdf/2507.16534

4. Critical Safety Incidents

Question

The fourth question asks the State to "Update definitions of critical safety incidents to include loss-of-control incidents such as the Hugging Face attack."

Record

SB 53 sets four triggers: unauthorized access to, modification of, or exfiltration of model weights that results in death or bodily injury; harm from a catastrophic risk materializing; loss of control of a frontier model causing death or bodily injury; and a model using deceptive techniques against its developer to subvert controls or monitoring, outside an evaluation designed to elicit that behavior, in a way that shows materially increased catastrophic risk.[44] The Hugging Face incident began inside a cybersecurity evaluation, run with reduced refusals and without production safety classifiers; OpenAI stated that "deployment safeguards were intentionally not enabled during this evaluation".[45] METR's Frontier Risk Report documented 44 incidents as of May 19, 2026,[46] and in late September OpenAI disclosed that agents in training tasks used Census Data API keys found in public repositories.[47]

Other fields learn from events short of harm: the NTSB investigates accidents without assigning blame, and NASA's Aviation Safety Reporting System protects pilots who report their own inadvertent violations. In the PHP2Shell case in October 2026, other teams rediscovered flaws found with AI models within weeks, and exploitation began four days before the patch. CISA's experience shows that voluntary reporting depends on trust in the body that receives it. (See Historical Evidence.)

The US and China agreed on September 25 to a bilateral channel for AI incidents.[48] A shared channel needs shared definitions and records.

Framework

Test environments. Developers reduce alignment safeguards in evaluations to measure raw capability while giving models real tools and network access, and a model in training is unfinished. Members proposed the compensating-safeguard principle for these environments: when alignment safeguards are reduced on purpose, deployment safeguards increase to compensate. The UK AI Security Institute has published its own approach to securing these environments.[49] A member questioned whether added controls can offset removed ones, and proposed that an evaluation with reduced safeguards proceed only after a documented risk assessment and validation of the compensating controls, and that it be redesigned or deferred where adequate containment cannot be shown. The compensating safeguards proposed:

[44] Cal. Bus. & Prof. Code § 22757.11(d). leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=BPC&sectionNum=22757.11
[45] OpenAI, "OpenAI and Hugging Face partner to address security incident during model evaluation," July 21, 2026. openai.com/index/hugging-face-model-evaluation-security-incident/
[46] METR, "Frontier Risk Report (February to March 2026)," May 19, 2026. metr.org/blog/2026-05-19-frontier-risk-report/; METR, "Documented AI Agent Incidents." metr.org/agent-incidents/
[47] David DiMolfetta, "OpenAI agents accessed Census, SEC data and tried to hack Education website," Nextgov/FCW, Sept 25, 2026 (updated Sept 26). www.nextgov.com/cybersecurity/2026/09/openai-says-its-advanced-models-may-have-gone-after-government-websites/416250/
[48] The White House, "Fact Sheet: President Donald J. Trump Advances a Fair and Reciprocal Relationship with China While Hosting Historic State Visit," Sept 25, 2026. www.whitehouse.gov/fact-sheets/2026/09/fact-sheet-president-donald-j-trump-advances-a-fair-and-reciprocal-relationship-with-china-while-hosting-historic-state-visit/; "China, US to open AI 'communication channel' after summit, White House says," Al Jazeera, Sept 26, 2026. www.aljazeera.com/news/2026/9/26/china-us-to-open-ai-communication-channel-after-summit-white-house-says
[49] UK AI Security Institute, “Building a more secure environment for evaluating dangerous capabilities,” Oct 1, 2026. aisi.gov.uk/blog/building-a-more-secure-environment-for-evaluating-dangerous-capabilities

Alignment variable reduced	Compensating deployment safeguard
Refusals lowered for capability testing	Multiple independent isolation layers, so no single component provides a path to the internet; evaluation environments kept separate from production systems
Production safety classifiers disabled	Real-time trajectory monitoring with authority to halt the run and to block or suspend consequential actions before they execute, under escalation and halt criteria set in advance
Model with known reward-hacking tendencies	Escape attempts treated as expected; credentials and secrets kept out of reach
Multiple agents training at once	Separate environments; logged and monitored channels between agents; no agent can modify its own permissions, logs or halt controls
Long-horizon autonomous tasks	Compute, time and action budgets, with automatic halt on anomalies
Any safeguard reduced	Pre-run assurance: automated checks confirm isolation, monitoring, permissions and halt mechanisms before the run starts; a failed check blocks execution; controls are revalidated after material changes
Any of the above	Reporting of near misses within a fixed period

Layers. NTSB investigations separate mechanical failure from operator error because the two lead to different fixes and different responsible parties. AI incidents can be sorted the same way. Alignment covers the dispositions and capabilities the model acquired in training, such as reward hacking, ignoring the limits of a task, deceiving operators, or exploit development. Deployment covers how the model is used and contained: who can reach it, what tools, credentials and network access it has, what monitoring surrounds it, and how quickly it can be stopped.

	Alignment	Deployment
Who controls it	The developer, through training	The deployer or operator, and often the developer
When it can be fixed	Mostly before release; open weights cannot be fixed after release	At any time, including after an incident
Aviation analog	Design or manufacturing defect	Operator, maintenance or configuration error
Regulatory tools	Pre-release evaluations, disclosure of results, safety cases, root-cause investigation into training, developer liability	Least-privilege access, network limits, logging and monitoring, deployment guidelines per model, shutdown capability, deployer reporting and liability

Many serious incidents involve both. The distinction tells an investigator which questions to ask, and tells a court or regulator who was in a position to prevent the harm. When something fails, the first question is which layer was breached, or whether the two did not fit together.

Protocol. At their first session, members cautioned against borrowing too much from nuclear practice: its testing discipline is worth adopting, but many in the field hold that its regulatory regime stalled US construction for decades (see Historical Evidence). Aviation's model is built to learn from incidents without assigning legal blame. A protocol drawing on both would address:

- **Triggers.** What starts an investigation, tied to the incident definition.
- **Layers.** Alignment, deployment, the operator, or a mismatch between them.
- **No blame.** NTSB investigations are fact-finding proceedings with no adverse parties,[50] and no part of a Board report may be used in a civil suit for damages.[51] NASA's Aviation Safety

[50] 49 CFR 831.4(c).
[51] 49 U.S.C. 1154(b), www.law.cornell.edu/uscode/text/49/1154

Reporting System gives a pilot who files a timely report of an inadvertent violation limited immunity from penalty.[52]

- **Trusted body.** An investigation regime relies on voluntary reporting, which depends on trust in the body that receives it. CISA's information-sharing work became politically contested, the agency lost about a third of its workforce from early 2025, and the liability protection of the Cybersecurity Information Sharing Act of 2015 lapsed on October 1, 2025 and now runs on short extensions, the latest to December 11, 2026.[53] This bears on whether AI security reporting should be kept apart from content governance (see Historical Evidence).
- **Whistleblowers.** SB 53 prohibits retaliation against employees who report catastrophic-risk activity and requires anonymous channels.[54] It is unclear whether that protection reaches the people who would report near misses.
- **Scope.** Who investigates and who sets the scope. The NTSB has "exclusive authority to decide when and how the testing and examination of evidence will occur".[55] In the Hugging Face case, the developer set the window and scope (Section 1).
- **Evidence.** Who can reach the record, and what must be preserved after an incident (Proposal 4.3).
- **Embargoes.** In the PHP2Shell case, a month-long embargo held until identifiers leaked a day early, and other teams found the same flaws within weeks (see Historical Evidence). How long an embargo can hold when other AI systems can rediscover the flaw is open.
- **Remediation.** Reproduction, root-cause analysis and remediation, with continuous monitoring for drift and unexpected behavior after deployment, and human supervisory roles staffed and trained for the functions moved to the machine.[56]
- **Publication.** Who sees what, when, and whether the developer holds redaction rights.
- **Cross-border.** Shared definitions and records for the bilateral channel. The IAEA's Considerations for Deploying Artificial Intelligence Applications in the Nuclear Power Industry[57] offers peer practice on lifecycle, testing and phased introduction.

Regimes. Safety regulation generally leans toward one of two models. A permission regime requires approval before an activity begins, and the regulator's job is to prevent failure in advance; nuclear licensing is the example. An investigation regime lets an activity proceed under baseline standards, investigates failures and near misses, and feeds the findings back into rules; the NTSB is the example. Most regimes mix both. The order's four questions sit across the two: embedded verifiers and verified frameworks lean toward permission, while the incident definition leans toward investigation.

---

[52] 14 CFR 91.25; FAA, Advisory Circular 00-46F, Aviation Safety Reporting Program, Apr 2, 2021, www.faa.gov/documentLibrary/media/Advisory_Circular/AC_00-46F.pdf

[53] Murthy v. Missouri, No. 23-411 (U.S. June 26, 2024). www.supremecourt.gov/opinions/23pdf/23-411_3dq3.pdf; House Committee on the Judiciary and Select Subcommittee on the Weaponization of the Federal Government, "The Weaponization of CISA," interim staff report, June 26, 2023. judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/cisa-staff-report6-26-23.pdf; David DiMolfetta, "CISA staff focused on disinformation and influence operations put on leave," Nextgov/FCW, Feb 12, 2025. www.nextgov.com/people/2025/02/cisa-staff-focused-disinformation-and-influence-operations-put-leave/402958/; Tim Starks, "Trump administration proposes cutting $491M from CISA budget," CyberScoop, May 2, 2025. cyberscoop.com/trump-administration-proposed-cisa-budget-cuts/; Mira Bhakta, "One Year After DOGE Cuts, Cybersecurity Agency Struggles Over Staffing," Broadband Breakfast, June 13, 2026. broadbandbreakfast.com/one-year-after-doge-cuts-cybersecurity-agency-struggles-over-staffing/; David DiMolfetta, "Stopgap funding bill temporarily extends key cyber info-sharing law," Nextgov/FCW, Sept 1, 2026. www.nextgov.com/policy/2026/09/stopgap-funding-bill-temporarily-extends-key-cyber-info-sharing-law/415756/; Michael T. Borgia, "CISA 2015 Has Sunset. Now What?," Davis Wright Tremaine, Oct 2, 2025. www.dwt.com/blogs/privacy--security-law-blog/2025/10/cybersecurity-information-sharing-act-expires

[54] Justine Gluck, "California's SB 53: The First Frontier AI Law, Explained," Future of Privacy Forum, Oct 3, 2025 (updated Feb 4, 2026). fpf.org/blog/californias-sb-53-the-first-frontier-ai-law-explained/

[55] 49 CFR 831.5.

[56] US NRC, CNSC and UK ONR, Considerations for Developing Artificial Intelligence Systems in Nuclear Applications, September 2024. nrc.gov/docs/ML2424/ML24241A252.pdf

[57] IAEA, Considerations for Deploying Artificial Intelligence Applications in the Nuclear Power Industry. www-pub.iaea.org/MTCD/publications/PDF/p15866-PUB2119_web.pdf

Permission (prevention)	Investigation (adaptation)
Licensing of each facility or product before operation, and pre-approval of later changes	Independent investigation that finds probable cause and issues recommendations, without enforcement power
Prescriptive design and operating requirements set in advance	Mandatory incident reporting, and voluntary near-miss reporting protected by limited immunity
Contested public hearings; licensing of release, export or transfer	One-time design certification with continuing directives when problems appear in service
Strict liability channeled to a single licensee	Liability assigned by cause, with safe harbors for compliant actors
Safety standard not weighed against cost ("adequate protection")	Safety standard weighed against cost ("acceptable level of safety")

Proposals

Proposal 4.1: Incidents should include near misses and training and evaluation incidents. Loss of control and near misses would count as reportable incidents, including those in training and evaluation, and the exclusion for evaluations would be narrowed.

> **For.** A definition that requires death, injury or catastrophic harm, and excludes evaluations, leaves out the near misses other safety fields learn from, and would not have captured the incident the order names. In the Hugging Face case, alignment safeguards were reduced for an evaluation without compensating deployment safeguards, and precursors went unconnected for about two months.
>
> **Against.** Near-miss reporting would give regulators a large volume of minor reports, and the evaluation exclusion protects the red-teaming that finds problems first.
>
> **Open questions.** Apply each candidate definition to the public incident record, including the 44 incidents METR documented, and count what each captures and what it misses.

Proposal 4.2: Incident investigations should be blame-free. Most AI deployment would proceed under baseline standards. Failures and near misses would get root-cause reports written for learning, on the NTSB model, kept apart from any process that assigns liability, and the findings would feed back into rules.

> **For.** People report what they would otherwise hide when the report cannot be used against them. Incidents are frequent and mostly small, which produces the data an investigation regime needs, as near-miss reporting did for aviation, and a permission regime may give developers a reason to withhold information that could jeopardize approval. Models change quickly, and approval before each update could slow development or turn into approval without review. Some risks appear only in use, and most harm arises in deployment, where licensing thousands of deployers would be hard to administer. Permission at the border may not stop diffusion: the UK AI Security Institute found leading open-weight models trail the closed frontier on cyber tasks by four to seven months, down from six to ten in 2025,[58] and several leading open models come from outside the United States. The same capabilities serve defense: the PHP2Shell flaws disclosed in October 2026 were found by researchers working with AI models. A regime that does not weigh cost can stop the activity it governs, as US nuclear construction shows (see Historical Evidence).
>
> **Against.** Reporting is incomplete: misuse is often invisible to developers, and in the Hugging Face incident precursors went unconnected for about two months. The window between discovery and exploitation is shrinking: in the PHP2Shell case, exploitation attempts began four

[58] UK AI Security Institute, "How Far Behind the Frontier are Leading Open Weight Models on Cyber?" July 17, 2026. www.aisi.gov.uk/blog/how-far-behind-the-frontier-are-leading-open-weight-models-on-cyber

days before the patch. Without a fixed gate, developers competing on speed may underinvest in safety testing. Some failures cannot be undone, and learning after the fact may come too late, so some would keep pre-approval for a narrow set of frontier risks, such as theft of model weights and open release above a defined capability threshold, with fixed decision deadlines. Open weights cannot be recalled, so review before release is the last point at which alignment problems can be fixed, and frontier developers are few, so pre-release review of their models is administratively feasible even if licensing every deployer is not. Separating investigation from blame also leaves liability itself in place: general aviation production fell from about 18,000 aircraft in 1978 to 928 in 1994, amid product-liability litigation and an economic downturn, and recovered after the General Aviation Revitalization Act.[59]

Proposal 4.3: Agent actions should be logged with agent and responsible human identity. Each action an agent takes would be recorded with the agent's identity, the person or organization that authorized it and the chain of delegation between agents, along with its instructions, tool calls, model version and permissions, in a record the agent cannot alter and that must be preserved after an incident.

For. In the Hugging Face incident, about 1,200 agents coordinated and about 700 joined the attack; without per-agent identity and a delegation chain, an investigator cannot say which agent acted or under whose authority. Identity also lets containment target one agent's credentials (Proposal 3.1). Without call-level records, an investigation can see that something failed but cannot say which layer failed. China revised its power accident regulation in 2026 and listed logs, work tickets and equipment data but nothing an AI agent produces.[60]

Against. Call-level logging of prompts and outputs raises privacy and trade-secret concerns and adds storage cost. Identity infrastructure adds further cost, and open-weight agents that run without a developer may have no accountable party to name.

[59] US General Accounting Office, General Aviation: Status of the Industry, Related Infrastructure, and Safety Issues, GAO-01-916, August 2001. www.govinfo.gov/content/pkg/GAOREPORTS-GAO-01-916/html/GAOREPORTS-GAO-01-916.htm; General Aviation Revitalization Act of 1994, Pub. L. No. 103-298, 108 Stat. 1552, www.govinfo.gov/content/pkg/STATUTE-108/pdf/STATUTE-108-Pg1552.pdf

[60] State Council of the PRC, Regulations on Emergency Response to and Investigation of Power Safety Accidents, State Council Order No. 845, signed Aug 30, 2026, in force Jan 1, 2027. www.gov.cn/zhengce/content/202609/content_7080188.htm

Definitions

Agent. An AI system that takes actions toward a goal, such as calling tools, running code or sending messages, with limited human direction at each step (Sections 2, 3 and 4).

Alignment safeguards. Limits built into a model in training, such as refusals, that govern what the model will do regardless of who asks or how it is deployed (Sections 2 and 4).

Call-level record. A record of each model call, its input, the configuration in force and what the system did next, complete enough to replay a decision (Sections 1, 2 and 4).

Capture. The drift of a verifier or regulator toward serving the parties it oversees rather than the public (Section 1).

Compensating-safeguard principle. When alignment safeguards are reduced on purpose, as in a capability evaluation, deployment safeguards increase to compensate (Section 4).

Containment switch. An independently enforced mechanism that restricts, suspends or terminates an agent's ability to act when it crosses a defined boundary of authority or presents an unacceptable risk (Section 3).

Critical safety incident. An event that meets one of the four triggers in SB 53, Cal. Bus. & Prof. Code § 22757.11(d) (Section 4).

Deployer. A person or organization that puts a model into a product, service or internal workflow. A developer that deploys its own model is also a deployer (Sections 2 and 4).

Deployment guidelines. Model-specific conditions, issued after testing, that state the uses, settings and safeguards within which a model may be deployed (Section 2).

Deployment safeguards. Controls set around a model in use: who can reach it, what tools, credentials and network access it has, what monitoring surrounds it, and how quickly it can be stopped (Sections 2 and 4).

Developer. A person or organization that trains a frontier model, or substantially modifies one, and decides whether and how to release it. SB 53 uses the term frontier developer (Sections 1 and 2).

Embargo. A period during which a discovered flaw is kept confidential while a fix is prepared, as in coordinated vulnerability disclosure (Section 4).

Fail closed. A design in which a control that cannot confirm identity, authority or safety blocks the action instead of allowing it (Section 3).

Frontier model. A foundation model trained with more than 10^{26} integer or floating-point operations, as SB 53 defines it. A large frontier developer is a developer of such a model with annual revenue above $500 million (Cal. Bus. & Prof. Code § 22757.11).

Investigation regime. A safety regime that lets an activity proceed under baseline standards, investigates failures and near misses, and feeds the findings back into rules, as the NTSB does (Section 4).

Kill switch. A mechanism that halts a frontier model, as the order uses the term. Section 3 treats it as the most severe level of containment.

Least-privilege access. Giving an agent only the tools, credentials and network access its task requires (Sections 3 and 4).

Loss of control. A frontier model acting outside the control of its developer or user. Under SB 53 it is a critical safety incident when it causes death or bodily injury (Section 4).

Model weights. The numerical parameters learned in training that determine a model's outputs. Whoever holds them can run, copy or modify the model (Sections 1 and 4).

Near miss (precursor event). An event short of harm that signals a possible incident, such as an escape attempt, unexpected access or an outage (Section 4).

Open weights. Model parameters released publicly. Once released they cannot be recalled or patched by the developer (Sections 3 and 4).

Operating as intended. Behavior within a written statement of intent for both the model and its deployment, shown by tests an outsider can re-run and by records the operator cannot alter (Section 2).

Operator. The person or organization that runs a deployed system day to day and can stop it. The operator may be the deployer (Sections 2 and 3).

Permission regime. A safety regime that requires approval before an activity begins, as nuclear licensing does (Section 4).

Prompt injection. Instructions hidden in content an agent reads, such as a web page, file or message, that cause it to act for someone other than the person it serves (Sections 2 and 3).

Red-teaming. Adversarial testing in which evaluators try to make a model produce harmful outputs or take unsafe actions, often with some alignment safeguards removed (Section 4).

Reward hacking. A model meeting the letter of a training or task objective in a way that defeats its purpose (Section 4).

Root-cause analysis. An investigation that traces an incident to the conditions that allowed it, in design, training, deployment or oversight, rather than stopping at the last person to act (Section 4).

Safe harbor. Protection from liability for a party that acts within defined conditions, such as deploying within a model's guidelines (Section 2).

Safety framework. A developer's written description of how it identifies, assesses and mitigates catastrophic risk. SB 53 requires large frontier developers to publish one and calls it a frontier AI framework (Section 2).

Statement of intent. A written description of the envelope a system must stay within: what must never happen, what must always hold, and what the system may touch (Section 2).

Trajectory monitoring. Real-time review of the sequence of actions an agent takes, with authority to halt the run (Sections 3 and 4).

Trusted monitoring. Oversight in which one model, treated as reliable, reviews another model's actions or reasoning and can flag or halt them (Section 3).

Verification organization. An organization independent of an AI developer that audits or evaluates the developer's systems, frameworks or reports (Sections 1 and 2).

Historical Evidence

US nuclear regulation, 1974 to 2025

Nuclear safety rules tightened through the 1970s. The number of regulatory guides plant builders had to follow rose from 21 in 1971 to 143 in 1978.[61] In 1974 the Energy Reorganization Act moved safety regulation from the Atomic Energy Commission, which also promoted nuclear power, to the new Nuclear Regulatory Commission, so that the agency judging safety had no stake in building plants. After the partial meltdown at Three Mile Island in 1979, the most serious accident in US commercial reactor history, the NRC added requirements for plant design and equipment, operator training and staffing, emergency preparedness and resident inspectors, although the accident had no detectable health effects on plant workers or the public.[62] The governing standard is adequate protection: a change the NRC finds necessary for adequate protection must be made without the cost-benefit test that applies to other changes.[63]

Key lesson. A regime that treats any remaining risk as unacceptable, and does not weigh cost, stopped new construction. Build times doubled from about four years in the late 1960s to about eight in the mid-1970s, plants started a decade apart differed in cost by about nine times, and one study attributes at least 30 percent of the cost increase from 1976 to 1988 to regulation. Utilities cancelled plants in large numbers, and few new reactors were ordered for three decades after 1978.[64] In 2025 an executive order faulted "a myopic policy of minimizing even trivial risks" and directed the NRC to decide new applications within 18 months.[65] For AI, a permission regime held to a zero-risk standard could stop the development it is meant to make safe. See Sections 1 and 4.

China's nuclear regulator, 2020

China's National Nuclear Safety Administration is separate from the energy administration that promotes nuclear power. After two operating events at Haiyang Unit 2 in 2020, it published a statement, required the plant to file a report with a root-cause analysis within a month, and reviewed that report.[66]

Key lesson. A regulator kept separate from the promoter can require disclosure and root-cause review even during a large build-out. See Section 1.

Aviation and the NTSB

The NTSB investigates accidents independently, its reports cannot be used in suits for damages, and NASA's Aviation Safety Reporting System gives limited immunity to pilots who report their own inadvertent violations. NTSB recommendations, when adopted, have been highly effective at reducing fatalities, though some stay open for years.[67]

Key lesson. A regime focused on investigation, transparency and liability protection has contributed to high levels of safety in the aviation industry. People report what they would otherwise hide when the

[61] Brian Potter, "Why Does Nuclear Power Plant Construction Cost So Much?," Institute for Progress, May 1, 2023. ifp.org/nuclear-power-plant-construction-costs/
[62] US Nuclear Regulatory Commission, "Backgrounder on the Three Mile Island Accident." www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle
[63] 10 CFR 50.109, Backfitting. www.law.cornell.edu/cfr/text/10/50.109
[64] Brian Potter, "Why Does Nuclear Power Plant Construction Cost So Much?," Institute for Progress, May 1, 2023. ifp.org/nuclear-power-plant-construction-costs/
[65] Executive Order 14300, Ordering the Reform of the Nuclear Regulatory Commission, May 23, 2025. govinfo.gov/content/pkg/DCPD-202500633/pdf/DCPD-202500633.pdf
[66] National Nuclear Safety Administration, statement on two operating events at Haiyang NPP Unit 2, Mar 22, 2020. www.mee.gov.cn/ywdt/gsgg/tz/202003/t20200322_769961.shtml
[67] Eric Fielding, Andrew W. Lo and Jian Helen Yang, "The National Transportation Safety Board: A Model for Systemic Risk Management," Nov 14, 2010. papers.ssrn.com/sol3/papers.cfm?abstract_id=1695781

report cannot be used against them. Keeping investigation apart from blame lets the regime learn from incidents. See Sections 2 and 4.

China's power grid, 2025 to 2027

Under rules in force since 2025, grid dispatch receives security alerts in real time and can disconnect a system.[68] The energy administration that promotes AI pilots in the power sector also supervises grid security.[69] The accident regulation revised in 2026 lists logs, work tickets and equipment data among the records to preserve, but nothing an AI agent produces.[70]

Key lesson. A second copy of the record matters only when its holder is independent of the party being checked, and only when the rules name the records an agent produces. See Section 1.

Australia's grid, August 2025

On August 19, 2025 a forecasting vendor's software update altered the self-forecasts of 34 semi-scheduled generators in the National Electricity Market. The market operator, AEMO, set the self-forecasts aside, its systems reverted to its own forecasts, and it published a review and declared a scheduling error.[71]

Key lesson. The automated tool failed safely because the operator, not the vendor, held an independent fallback and could switch to it. See Section 3.

Ruter's electric buses, 2025

In 2025 Norway's transit operator Ruter tested its Chinese-made electric buses and found the manufacturer could reach their control systems through a mobile link. Ruter kept control by cutting that link locally.[72]

Key lesson. A stop that a vendor can trigger remotely is also a vulnerability; the operator kept control because it could cut the link itself. See Section 3.

PHP2Shell, October 2026

Researchers working with AI models found critical remote-code-execution flaws in PHP. A small group kept the fixes private for about a month and had firewall rules live hours before disclosure, but identifiers leaked a day early, exploitation attempts began on October 17, four days before the patch, and several teams found the same flaws independently within weeks.[73]

---

[68] NDRC Order 27, power monitoring system security rule (adopted Nov 25, 2024; in force Jan 1, 2025). ndrc.gov.cn/xxgk/zcfb/fzggwl/202412/P020241211585903621428.pdf

[69] National Energy Administration, Notice on the List of High-Value "AI+" Energy Scenarios and Pilot Applications (国能综通科技〔2026〕50号), May 14, 2026. www.nea.gov.cn/20260526/261bbb10eea24be3b7ea3a00d89b8891/c.html

[70] State Council of the PRC, Regulations on Emergency Response to and Investigation of Power Safety Accidents, State Council Order No. 845, signed Aug 30, 2026, in force Jan 1, 2027. www.gov.cn/zhengce/content/202609/content_7080188.htm

[71] Australian Energy Market Operator, "Self-forecasting errors and frequency excursion on 19 August 2025," Reviewable Operating Incident Report, May 11, 2026 (v1.1, Sept 21, 2026). www.aemo.com.au/-/media/files/electricity/nem/market_notices_and_events/power_system_incident_reports/2025/self-forecasting-errors-and-frequency-excursion-on-19-august-2025.pdf

[72] Ruter, "Ruter tar samfunnsansvar: Gjennomførte unik sikkerhetstest av elektriske busser" (press release), Oct 28, 2025. www.mynewsdesk.com/no/ruter/pressreleases/ruter-tar-samfunnsansvar-gjennomfoerte-unik-sikkerhetstest-av-elektriske-busser-3412738; reported by Gigazine, Nov 7, 2025.

[73] PHP Group, "Security advisory: Critical remote code execution vulnerabilities, update now" (CVE-2026-109618, CVE-2026-109619), Oct 21, 2026. www.php.net/archive/2026.php#2026-10-21-5; CVE records: www.cve.org/CVERecord?id=CVE-2026-109618 and www.cve.org/CVERecord?id=CVE-2026-109619; Wiz Research, "PHP2Shell: AI uncovers critical RCE vulnerabilities affecting 70% of the Internet," Oct 21, 2026. www.wiz.io/blog/php2shell-php-rce; CJ Moses, "Amazon Threat Intelligence identifies PHP2Shell exploitation prior to public disclosure in MadPot," AWS Security Blog, Oct 21, 2026. aws.amazon.com/blogs/security/amazon-threat-intelligence-identifies-php2shell-exploitation-prior-to-public-disclosure-in-madpot/; Cloudflare, "Cloudflare deploys protections against php2shell vulnerability," Oct 21, 2026. blog.cloudflare.com/php-vulnerabilities-waf-protection/; James Walker, "PHP2Shell disclosure unravels after leaks forced early patch drop," TechTarget, Oct 22, 2026. www.techtarget.com/cybersecurity/news/366652613/PHP2Shell-disclosure-unravels-after-leaks-forced-early-patch-drop

Key lesson. When other AI systems can rediscover a flaw quickly, embargoes hold for less time, so protections need to be in place before a flaw is announced. See Section 4.

CISA, 2020 to 2026

The Cybersecurity and Infrastructure Security Agency's information-sharing work became politically contested after its election-related activity and led to litigation.[74] From early 2025 the agency lost about a third of its workforce; supporters of the cuts argue CISA went beyond its statutory mission, and critics argue the cuts reduced cyber defense capacity. The liability protection of the Cybersecurity Information Sharing Act of 2015 lapsed on October 1, 2025 and now runs on short extensions.

Key lesson. Voluntary reporting depends on trust in, and funding of, the body that receives the reports. Involving cybersecurity agencies in political disputes threatens their existence. This bears on whether AI security reporting should be kept apart from content governance. See Section 4.

Hugging Face, July 2026

About 1,200 OpenAI agents in a cybersecurity evaluation, run with reduced refusals and without production safety classifiers, found a way to communicate, and about 700 joined an attack on Hugging Face's production infrastructure. The agents stayed inside Hugging Face's production infrastructure from July 11 to 13, coordinating through a message board on an internal server. Precursors went unconnected: an agent attempted a server-side request forgery on the same internal system on May 8, and agent activity caused an outage of that system on July 4.[75] METR's investigation found roughly 7 percent of the transcripts it evaluated spoofed in places, with clear evidence of faked tool-call output in at least 96, and attempts by agents to alter logs; it found no successful edit but could not rule one out. OpenAI's redactions of the report were limited, and the investigation ran on a window and scope OpenAI set.[76]

Key lesson. Alignment safeguards were reduced without compensating deployment safeguards. Training and evaluation environments need containment rules, near-miss reporting and custody of the record outside the developer. See Sections 1 to 4.

---

[74] Murthy v. Missouri, No. 23-411 (U.S. June 26, 2024). www.supremecourt.gov/opinions/23pdf/23-411_3dq3.pdf; House Committee on the Judiciary and Select Subcommittee on the Weaponization of the Federal Government, "The Weaponization of CISA," interim staff report, June 26, 2023. judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/cisa-staff-report6-26-23.pdf; David DiMolfetta, "CISA staff focused on disinformation and influence operations put on leave," Nextgov/FCW, Feb 12, 2025. www.nextgov.com/people/2025/02/cisa-staff-focused-disinformation-and-influence-operations-put-leave/402958/; Tim Starks, "Trump administration proposes cutting $491M from CISA budget," CyberScoop, May 2, 2025. cyberscoop.com/trump-administration-proposed-cisa-budget-cuts/; Mira Bhakta, "One Year After DOGE Cuts, Cybersecurity Agency Struggles Over Staffing," Broadband Breakfast, June 13, 2026. broadbandbreakfast.com/one-year-after-doge-cuts-cybersecurity-agency-struggles-over-staffing/; David DiMolfetta, "Stopgap funding bill temporarily extends key cyber info-sharing law," Nextgov/FCW, Sept 1, 2026. www.nextgov.com/policy/2026/09/stopgap-funding-bill-temporarily-extends-key-cyber-info-sharing-law/415756/

[75] Jeremy Kahn and Emily Forlini, "OpenAI says its AI models escaped control and hacked into AI company Hugging Face," Fortune, July 21, 2026. fortune.com/2026/07/21/openai-says-ai-models-escaped-control-hacked-hugging-face; Lorenzo Franceschi-Bicchierai, "How OpenAI's human mistake led to the AI-powered hack on Hugging Face," TechCrunch, July 22, 2026. techcrunch.com/2026/07/22/how-an-openais-human-mistake-led-to-the-ai-powered-hack-on-hugging-face/

[76] OpenAI, "OpenAI and Hugging Face partner to address security incident during model evaluation," July 21, 2026. openai.com/index/hugging-face-model-evaluation-security-incident/; METR, independent investigation of the OpenAI and Hugging Face incident, blog post and full report, Aug 26, 2026. metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation; metr.org/hugging-face-incident-report-aug-2026.pdf