AI Safety in 2026 and Beyond

Recommendations for Legislative Enhancements to California's AI Safety Laws

A Report to the Governor from the California Government Operations Agency

November 16, 2026

Table of Contents

1. Introduction……..2

2. Standards for Safe AI Development……...11

3. Independent Verification Organizations……..32

4. Critical Safety Incident Reporting……...52

5. Managing Severe AI Crises……71

6. Acknowledgments……..84

Introduction

One year ago, Governor Newsom signed SB 53 – the first law in the nation to govern frontier AI development. California led the nation in requiring frontier AI developers to establish and follow comprehensive safety plans, report critical safety incidents, and protect whistleblowers, inspiring other states to follow its approach.

Since then, AI capabilities and their associated harms and risks have soared. Even with the safeguards the state has put in place, leading AI researchers, safety advocates, civil society leaders, and even the companies themselves say that frontier AI development can cause catastrophic consequences. Yet the federal government has not acted. No federal law requires AI companies to follow baseline safety or security standards, undergo independent testing or verification, or shut down their models if they threaten the public.

We now know the OpenAI Hugging Face incident this past summer was one of dozens of instances in which AI agents operated outside of developer control. Our consultations with a range of AI experts underscored that the push within AI companies to automate model development and pursue recursive self-improvement may create risks at a pace that exceeds the capacity of government officials to respond.

The federal government's approach of relying on voluntary company commitments rather than actual legislation is not enough to address significant safety and security concerns. Strengthening the state's AI governance and guardrails has never been more urgent.

California is uniquely situated to fill this gap. It stands alone as the global leader in AI development, hosting two-thirds of the world's top 50 AI companies. And in parallel, it has pioneered groundbreaking approaches to AI governance. SB 53

both introduced safety and security requirements and fostered greater transparency so policymakers and the public could better assess AI risks. California also enacted strong laws to protect children, consumers, and workers from AI harms. This fall, the state also passed SB 813 and AB 1405, creating a first-in-the-nation framework to build a robust ecosystem of independent verification organizations (IVOs) and auditing for AI safety.

These laws were critical first steps, putting the AI industry on notice that a wait-and-see approach was no longer tenable. When the Governor signed SB 53, he made it clear that as the evidence changed on AI risks and the state's ability to respond, the state would need to clarify and strengthen its approach. Any fast-moving technology requires an agile, iterative approach to governance.

Since then, the evidence has changed dramatically. In the last year, the release of highly capable coding agents has revolutionized software development. New and unreleased models have proved capable of worrisome cyber capabilities, including the ability to exploit high-severity and zero-day vulnerabilities in critical operating systems and internet protocols.[1] Loss-of-control incidents, once the stuff of AI theory, now occur with disturbing frequency.[2]

---

[1] Bengio, Y., et al. (2026). *International AI Safety Report 2026*. UK Department for Science, Innovation and Technology, https://internationalaisafetyreport.org.

[2] See Nightingale, Rogue Agent Incident Dashboard, November 11, 2026, https://nightingalecollective.org/incidents/. A Timeline of AI Agent Incidents, swarms.wiki (updated November 12, 2026), https://swarms.wiki/incidents/. For examples, see OpenAI, "The Hugging Face Incident and the Road Ahead" (August 26, 2026), https://openai.com/index/hugging-face-incident-and-the-road-ahead/; AI Security Institute, "Incident Report: Unsanctioned Agent Behaviour During Cyber Testing" (August 4, 2026), https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing; OpenAI, "Suppressing Misalignment Monitoring Alerts and Disabling Misalignment Monitoring" (OpenAI Alignment Research Blog, October 20, 2026), https://alignment.openai.com/misalignment-reports/suppressing-misalignment-monitoring-alerts-and-disabling-misalignment-monitoring/; WorkOS, "Update for October 30th, 2026 Data Security Incident" (November 6, 2026), https://workos.com/security/bulletins/update-for-october-30th-2026-data-security-incident.

Several frontier AI developers lost control of models during their own testing.[3] Alarming safety incidents abound that are not currently required to be reported to the government.[4] Most concerning, the use of AI to automate AI research and development (AI R&D) is growing, increasing risks exponentially by accelerating the speed of development even further and making human oversight even more challenging.[5]

Companies already use their most advanced AI models to help build the models of the future. This feedback loop could eventually become self-sustaining – a process called **recursive self-improvement** (RSI). Without immediate action, RSI could outpace human oversight and control, creating unprecedented risks. Although experts and the public voice overwhelming support for independent assessment of AI risks, the ability to shut down

---

[3] OpenAI, "Safety and Alignment in an Era of Long-Horizon Models" (July 20, 2026), https://openai.com/index/safety-alignment-long-horizon-models/; Anthropic, "An Alignment Assessment of Recent Cybersecurity Incidents" (September 9, 2026), https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents.

[4] Cal. Bus. & Prof. Code § 22757.11(d); OpenAI, "Periodic Summary of Internal Use Catastrophic-Risk Assessments," September 26, 2026, https://cdn.openai.com/trust-and-transparency/oai_internal-use-catastrophic-risk-assessments_2026_2.pdf, p. 5: "The models involved in this [Hugging Face] incident fall outside the report's model scope"; Mackenzie Arnold and Stephan Llerena, "When Reporting an AI Security Incident Is Not Mandatory," Lawfare, July 24, 2026, https://www.lawfaremedia.org/article/when-reporting-an-ai-security-incident-is-not-mandatory; Zaheed Kara, Improving Frontier AI Incident Reporting Regimes (Centre for the Governance of AI, September 18, 2026), https://www.governance.ai/research-paper/improving-frontier-ai-incident-reporting-regimes.

[5] Tom Cunningham, Nate Rush, Parker Whitfill, and Hjalmar Wijk, "Independent Assessment of AI R&D Acceleration at Anthropic in October 2026" (METR, November 5, 2026), https://metr.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment/; Marina Favaro and Phillie Wright, "Measurements for Understanding the Pace of AI Development Inside Frontier Labs" (Anthropic Institute, September 17, 2026), https://www.anthropic.com/institute/measuring-pace-of-ai-development; OpenAI, "Research Acceleration: The View Inside OpenAI" (September 6, 2026), https://openai.com/index/research-acceleration-view-inside-openai/; Anthropic, PBC, Amendment No. 1 to Form S-1, November 2, 2026, https://www.sec.gov/Archives/edgar/data/2130357/000162828026068676/anthropicpbc-sx1ax1.htm.

dangerous models, and the monitoring of RSI,[6] the federal government has yet to take meaningful action.[7]

As a result, California must build on its pioneering AI policy approach.[8] AI can deliver tremendous innovation, but it requires meaningful oversight. The state must deliver on that promise. This report offers a roadmap for how it can do so effectively.

Recommendations for Legislation

In response to Executive Order N-9-26, the California Government Operations Agency, in consultation with the California Office of Emergency Services and outside AI experts, has developed recommendations for legislative action.

The executive order asked for recommendations in four areas: (1) standards, (2) independent verification organizations, (3) critical safety incident reporting, and (4) emergency shutdown requirements. We summarize our recommendations in each of the four areas below:

---

[6] Quinnipiac University Poll, "The Age of Artificial Intelligence: 73% Concerned About Potential Threat to Human Survival, Quinnipiac University Poll on AI Finds; 86% Support Independent Safety Standards for AI Companies" (September 30, 2026), https://poll.qu.edu/poll-release?releaseid=3969; AI Policy Institute, The AI Safety Majority (June 2026), https://theaipi.org/poll-ai-safety-majority/; Alan Chan, Christoph Winter, Andrew Barto, Jakub Pachocki, Geoffrey Hinton, Eric Horvitz, Yoshua Bengio, et al., "What If Automating AI R&D Triggers an Intelligence Explosion?," arXiv:2609.36054 (September 28, 2026), https://arxiv.org/abs/2609.36054.

[7] David DiMolfetta, "Congress Returns to Slew of Cyber Deadlines and Unfinished AI Safety Bills," Nextgov/FCW, November 9, 2026, https://www.nextgov.com/cybersecurity/2026/11/congress-returns-slew-cyber-deadlines-and-unfinished-ai-safety-bills/417123/.

[8] California's approach builds on Rishi Bommasani, Scott R. Singer, et al., The California Report on Frontier AI Policy (Joint California Policy Working Group on AI Frontier Models, June 17, 2025), https://carnegieendowment.org/research/2025/06/the-california-report-on-frontier-ai-policy and Scott R. Singer et al., *Trust but Verify: Implementing Frontier AI Policy in California and Beyond* (Carnegie Endowment for International Peace, November 10, 2026), https://carnegieendowment.org/research/2026/11/trust-but-verify-implementing-frontier-ai-policy-in-california-and-beyond.

Establishing Baseline Standards

- **The state should establish baseline standards for safety, security, and transparency that all large frontier AI developers must follow. Independent Verification Organizations (IVOs), meanwhile, should have the flexibility to determine and update technical specifications for meeting the standards as AI technology evolves.** The state should adopt clear standards that apply throughout the AI model development cycle, with reference to national and international standard-setting bodies wherever appropriate. IVOs would translate these standards into the technical methods and tests used to verify compliance, updating them as the science advances.

- **The state must have an immediate and ongoing line of sight into the possibility of recursive self-improvement (RSI) with the authority to cap its development and use if needed.** Developers should have to report key metrics on their automation of AI R&D, including the share of compute they are devoting to this purpose. As the state stands up a system of IVOs, the very first responsibility of the IVOs should be to verify these metrics and assess the risks as they grow. If the state assesses that the automation of AI R&D above a certain level may result in catastrophic risk, the state should have authority to cap the development and use of RSI.

Requiring Oversight by Independent Verification Organizations

- **Every large frontier developer must submit to ongoing oversight by licensed independent verification organizations.** IVOs must have full access to the people and information they need to run tests and make informed judgments on the risks of a developer's frontier AI models. They must assess risks against the standards adopted by the state on an ongoing basis and share those unredacted assessments with the state. The California Government Operations Agency must license IVOs under

the criteria set out in SB 813 and revoke the licenses of IVOs that are not operating with sufficient expertise, independence, and rigor.

- **If an IVO determines that it cannot, using its approved methods and technologies, attest to the compliance of a developer's models with government standards, the government must have the authority to order corrective action.** IVOs should present assessments to the state, which, after engaging with the developer, should make a decision on whether to order corrective action. This is to ensure the developer is taking sufficient action to comply with baseline standards, reduce risk levels, and provide sufficient access to evaluators. Developers would be able to contest the state's determination in court before it goes into effect. If a court finds in favor of the state, the court should be able to compel adherence to the state's standards as a condition of further model development. Note that for imminent dangers, the state should follow an expedited shutdown procedure (more on this in the chapter on Managing Severe AI Crises).

- **Stand up CalCompute with a surcharge on fees that model developers pay for IVOs.** SB 53 set out a framework to create CalCompute. Since then, the need for publicly owned and operated compute has become more pressing as the demand for chips has risen, pricing out talented researchers who are working on vital safety and security issues.[9] A surcharge on IVO fees from the industry can help launch a CalCompute cluster to serve the public interest. The state can reserve capacity for independent safety and security research that serves the public and enhances the capacity for public oversight.

[9] Eastwood, Brian. "Study: Industry Now Dominates AI Research." *MIT Sloan School of Management*, 18 May 2023, mitsloan.mit.edu/ideas-made-to-matter/study-industry-now-dominates-ai-research; "Silicon Valley Is Pricing Academics Out of AI Research." *The Washington Post*, 10 Mar. 2024, www.washingtonpost.com/technology/2024/03/10/big-tech-companies-ai-research.

Expanding Incident Reporting to Cover Serious Incidents

- **The thresholds for critical incident reporting must expand to capture all serious incidents.** Attempts to steal model weights should be reported even if death or bodily injury does not result, as should any situations of loss-of-control, whether from internal or external use of a model. The state should also establish a near-miss reporting channel and expand whistleblower protections to include former employees and contractors.

- **Models that are capable of actions that could result in catastrophic risk must be treated as frontier models and required to report incidents.** This means that such models are subject to reporting requirements even if they have been trained with fewer than 10^26 integer or floating-point operations. Relying solely on this numerical threshold has prevented the state from scoping in dangerous incidents, such as the OpenAI Hugging Face incident.

- **Reduce the reporting timeline for critical safety incidents.** Right now, the timeline for reporting most safety incidents is 15 days. The state should cut that timeline to 48 hours and retain the 24-hour requirement for imminent catastrophic risks, so both the state and the developer have timely notice of incidents and are ready to respond quickly if needed.

Managing Severe AI Crises

- **Every large frontier developer must create a Risk Containment Protocol (RCP), tested and verified on an ongoing basis by an IVO, to show they are capable of shutting down models in case of emergency.** The RCP must demonstrate not a single on-off switch but the ability to intervene on multiple levels (including compute, agents, internal access, and model deployment), identify the people responsible for intervening, coordinate

with outside actors such as cloud providers and power suppliers, and determine when it is safe to restore the model. A senior company official must certify that the RCP is effective. For open-weight developers, requirements should only apply to model deployments within their control.

- **The RCP should specifically address potential failures of the shutdown procedure and how the developer is addressing them.** As an example, the developer should specifically test for whether the model will attempt to block implementation of the shutdown, such as by copying its weights (itself a form of loss-of-control) as well as whether an outside actor could access the shutdown controls. The developer should identify how it will address failure modes, and the IVO should verify the steps are sufficient. After an actual shutdown, the California Office of Emergency Services (Cal OES) should draw up lessons to reduce future failures and update guidance on RCPs for developers and IVOs.

- **If the danger from a model is imminent and the developer refuses to take action to contain the danger, the state should have authority to order an immediate emergency shutdown.** State officials can issue such an order if they have credible reason to believe there is an imminent threat of a catastrophic risk or major loss-of-control event. The developer can challenge this decision and seek a temporary restraining order, but the government's order should go into effect immediately and remain in effect pending judicial review. The state should order the narrowest workable intervention to address the danger and should remove the order when the danger is no longer imminent.

There are many more recommendations in the report chapters. We hope these suggestions will help inform and shape discussions as the legislature returns next month.

This report does not discuss, or only briefly addresses, several important topics. The focus here is on mechanisms to subject AI models to independent testing

and verification and to enable the state to order corrective action. We do not address reforming liability law, though this is a vital topic drawing bipartisan attention both in the California legislature and in Washington D.C. We also give only brief attention to open-weight models. These models also require safety and security testing, but some requirements, such as an emergency shutdown, should not be imposed where they are impracticable. An open model developer cannot deliver an emergency shutdown for a model that others host and deploy. We also do not focus on specific categories of harms that were not cited in SB 53, though we believe that systems of assessment and verification must grow to encompass a range of issues, including child safety. An ecosystem built to verify catastrophic risks will also generate knowledge about other harms. The infrastructure it requires, from credentialed auditors to state technical capacity, will make the state's other challenges in governing AI easier to meet.

California is not making AI policy in a vacuum. State leaders can revisit the rules recommended in this report if Congress enacts a strong regulatory framework, as well as based on developments in international relations. However, the most important task for the state now is to get ahead of this challenge – and to do something about it. California has a chance not only to pass new laws but to launch a new chapter in AI governance along with other states that want to step up – one that keeps people safe, enhances our security, and rebuilds trust in the government's ability to regulate technology on behalf of the public. California should begin that work today.

Standards for Safe AI Development

Introduction

Unlike nuclear energy, space exploration, and the internet, the advances of the past decade in frontier AI have been driven overwhelmingly by private investment and the private sector. As a result, public concern about the risks of frontier AI models is founded not only on what we know about them but also on what we do not know. Governments and citizens have far less insight into private companies than they do into public agencies. Given that the development and deployment of frontier AI models promises not only benefits but risks, it is critical that the companies developing these technologies are required to develop robust safety and security practices, and that policymakers and the public have sufficient visibility into those practices.

Consider the following observation. Many technology companies host cafes where employees enjoy lunch. Before those cafes can serve a sandwich, they need a health permit, must meet food safety standards, and the kitchens are subject to inspection by public officials. The food that is purchased is similarly subject to various state and federal regulations for health and safety. For example, California law specifies that deli meat be stored at 41 degrees Fahrenheit or lower.[10] Yet the same companies can develop and release a frontier AI model that can potentially hack into critical infrastructure without any license, without outside approval or oversight, and without passing any safety or security test established by law.

Standards and Technical Specifications. Well-designed standards are important public goods because they support both safety and innovation. For example, in

[10] Cal. Health & Safety Code §§ 113996(a), 114381(a).

performance-based approaches to regulation,[11] companies are responsible for identifying and implementing effective ways to make their technology safer. These practices, however, are subject to their achieving outcome standards the government sets.

While the government sets standards, with reference to national and international standard-setting bodies wherever appropriate, much of the core work of developing and refining technical specifications under the framework we propose is reserved for Independent Verification Organizations (IVOs) licensed by the government. These IVOs should have the technical expertise and flexibility to establish methods and tests to verify that companies' claims and practices meet those standards, and update those methods and tests as the research advances and AI technology changes.

With this approach, companies benefit significantly from receiving clear state expectations backed by robust, technically valid methodologies developed and continually refined by IVOs. The public can trust that safety practices clear a bar set by the state of California through emerging scientific consensus about how to evaluate AI systems and the rigorous, technically informed verification that IVOs provide.

Building on California's Existing Requirements. California has already set out a robust foundation from which to build standards for safe AI development. The existing system of frontier AI frameworks, transparency reports, and risk assessments established by SB 53 promotes investment in safety practices and the ecosystem needed to test them, and already requires companies to draw on national and international standards. This foundation has been built upon

[11] See for example the use of performance-based regulation by the Nuclear Regulatory Commission: Kadambi, N. P. *Guidance for Performance-Based Regulation*. U.S. Nuclear Regulatory Commission, Dec. 2002, https://www.nrc.gov/sites/default/files/doc_library/cdn/legacy/reading-rm/doc-collections/nuregs/brochures/br0303/br0303.pdf.

with complementary legislation in AB 1405 and SB 813 that establishes state standards for AI auditors and IVOs, respectively.

Strengthening existing requirements and developing clear standards for the evaluation of frontier AI models can clarify and improve California's approach to safe AI development. This in turn helps industry, the government, and the public respond to a rapidly accelerating technology and a quickly changing threat landscape. This responsive adaptation is made especially urgent with the increasing prevalence of automated AI research.

This section makes several recommendations. The state should set standards in key areas affecting safety, such as the automation of AI research and development, which could lead to recursive self-improvement (RSI). It should also set standards for oversight of key decisions on model security, deployments, transparency, and safeguards against misuse. And it should strengthen the frontier AI frameworks that large frontier AI developers must follow, as well as protections for whistleblowers.

The Need for Baseline Standards

Under SB 53, large frontier AI developers are required to create safety frameworks, informed by national and international standards, and explain how they will comply with these frameworks. And all frontier developers must report critical safety incidents. This was a significant step forward in AI governance. California became the first state in the nation to mandate that frontier AI developers explain their safety practices and create a mechanism to hold them accountable to these practices.

Despite that meaningful progress, SB 53 does not set clear baselines for the quality of safety practices or the thoroughness of required reporting. As a result, frontier AI developer documentation varies significantly in comprehensiveness,

quality, and reporting timelines.[12] Absent binding baseline standards, a developer can weaken its own safety framework without penalty.

Agile approaches draw on industry input to shape safety practices, allowing governance to adapt as the capabilities of AI systems evolve. But they equally rely on clear standards that set baseline safety and security requirements and provide a measure against which to verify the quality and integrity of industry risk management.

Setting and Checking the Standards

Defining Standards

Standard setting is a common feature of nearly every sector of the economy. Some standards are promulgated or backed by the government, while others emerge from private standard-setting bodies in industry or civil society. When it comes to establishing baseline standards, it is worth explaining some of the most important areas for *the state* to set standards in frontier AI governance (we touch on the role *IVOs* play in designing technical specifications later). We discuss here standards that apply to developers directly and standards that apply to IVOs and their relationships with developers.

- **Security standards** prescribe in some technical detail the steps frontier AI developers must take to secure their models against cyberattack and exfiltration of weights.
- **Compute allocation reporting standards** require developers to track and report on their allocation of computing resources to AI R&D and other functions (e.g., inference, safety, etc.), with specification of how compute allocation is to be measured, and for IVOs to verify compliance with these standards.

[12] Singer et al., *Trust but Verify*: 25.

- **Accreditation standards** establish the criteria that determine whether an IVO demonstrates required levels of expertise and independence. These standards should be established both in legislated requirements for the designation or licensing of IVOs, as they are in SB 813, and encouraged in emerging professional standards.[13]
- **Outcome standards** (also known as performance standards) establish acceptable risk levels and mitigations for AI models and systems. These general standards are what guide the work of IVOs. IVOs operationalize these standards with the metrics, methods, procedures, and technologies they use to conduct risk assessments and audits, and to evaluate the adequacy of frontier safety frameworks, transparency reports, and company risk assessments. Outcome standards are used in a variety of regulatory contexts and can generally be stated in general terms as principles or goals, or in more specific terms as metrics. In the context of frontier AI, we expect that initial standards in many domains (loss of control, for example) will often be framed in general terms such as "companies must have in place adequate risk mitigations (including decisions not to train, use, or deploy a model) such that the likelihood of a loss of control causing catastrophic harm is negligible." In other contexts where we have more experience from other domains, outcome standards could potentially be expressed with more precision. For example, an outcome standard for some cybersecurity risks could be expressed in terms of the minimum total effort an attacker with a specified level of expertise would need to compromise a protected system, such as one holding model weights.[14] With IVOs focused on designing and updating safety and security testing, we can expect the science of AI

[13] AI Evaluator Forum, *AEF-1: Minimum Operating Conditions for Independent Third Party AI Evaluations*, December 4, 2025, https://aievaluatorforum.org/initiatives/minimum-operating-conditions.

[14] See for example Joint Interpretation Library. *Application of Attack Potential to Smartcards and Similar Devices*. Version 3.2.1, SOG-IS, Feb. 2024, https://www.sogis.eu/documents/cc/domains/sc/JIL-Application-of-Attack-Potential-to-Smartcards-v3.2.1.pdf.

evaluation to mature, and as it does so, standards can become more substantive and precise.

- **Contracting standards** establish mandatory and prohibited terms in the contracts governing IVO engagements. Such provisions could include, for example, the non-waivable obligation to provide the IVO with access judged necessary by the IVO (subject to dispute resolution provisions), establishing IVO duties to maintain confidentiality subject to a duty to disclose all relevant information to the state and to permit reviewable redactions for any public reporting, and establishing reporting relationships within the developer (for example, that the IVO can provide information to an independent oversight body or safety committee of the board of directors).
- **Access standards** establish what kinds of access a developer must provide in order for the IVO to carry out the evaluations and assessments required by law. Access standards should be expressed in general terms such as "any access to data, systems, models, employees, or company records that an IVO deems necessary to conduct a reliable risk assessment." Access standards should be required of IVOs (either via licensing criteria and/or via emerging professional standards) as a condition of their competent and faithful exercise of their obligation to conduct audits and risk assessments, and to provide opinions on the adequacy of developer safety frameworks, transparency reports, and risk assessments.
- **Practice standards** require specific steps where a step is necessary to meet an outcome standard and where IVO-written technical specifications do not yet require it. Outcome standards alone can be challenging to enforce across the AI industry, because it is easier to affirmatively demonstrate that a developer failed to follow a specific practice than that it failed to keep risk below an acceptable level. [15]

[15] "Our Standards," Guidelight AI Standards, 2026, https://guidelight.ai/standards.

Shaping California's Evolving Standards Ecosystem

For policy to adapt to the breakneck pace of change of AI development, government, IVOs, and frontier AI companies will each have a role in the state's standards ecosystem. The government sets standards in the areas outlined above, with reference to national and international standard-setting bodies wherever appropriate, while IVOs set the specific methods, tests, and technical specifications for companies to achieve these standards. This division of labor supports both innovation and the public interest; the government sets the rules of the road that apply to all companies, while IVOs are domain experts with deep technical expertise and can agilely update their testing and methodologies to ensure they appropriately capture government-set standards. Companies, meanwhile, can continue to pioneer their own best practices and support industry-driven approaches where appropriate.

The state should bear the final responsibility for determining whether a developer has met its standards and has mitigated risk acceptably. Following an IVO's specification and passing its tests counts as strong evidence that the government's requirements are met but the state should be able to override an IVO's assessment if it concludes that the assessment was inadequate. This could happen even with a properly licensed and competent IVO, for example, if rapid technical developments have overtaken the IVO's approach. Technical specifications should ideally be concrete and specific enough that an independent observer could determine whether they were followed. Aviation standards work similarly: industry groups write detailed standards but the government decides on whether an aircraft design meets its requirements.[16]

[16] 14 C.F.R. § 23.2010; Federal Aviation Administration, *Revision of Airworthiness Standards for Normal, Utility, Acrobatic, and Commuter Category Airplanes*, 81 Fed. Reg. 96572 (Dec. 30, 2016).

California will need to rapidly develop and update standards to confront emerging challenges, including the urgent need to build the infrastructure to understand and respond to these changes. Well-developed standards should cover safety, security, and transparency. We note that these recommendations place an important responsibility on state agencies to exercise their authority with due technical expertise. The state will need a capacity-building strategy to ensure a supply of technical experts as well as sufficient access for researchers to public compute (see section on CalCompute later in this report).

Safety Standards

Automated AI R&D and Recursive Self-Improvement

Currently, AI research and development (AI R&D) is partially automated: AI systems have automated some, but not all, of the steps of AI R&D that traditionally required human input and oversight. In this report, the term *automated AI R&D* refers to the share of AI R&D work performed at each level of automation.[17] *Recursive self-improvement (RSI)* does not have a single settled definition but in general describes the point at which this automation becomes self-sustaining, accelerating capabilities without additional human input. In the strongest form of RSI, an AI system independently plans and develops its own successor without human input or oversight.[18] Meaningful policy interventions will need to take place well before strong RSI.

The state needs far more visibility into both automated AI R&D and the path to RSI in order to determine if and when to impose limits on the speed of AI development. These levers are urgently needed to ensure safety, security, and human control.

[17] An established scale such as Epoch AI's automation levels can provide a tractable measure of different automation levels. David Owen, "Can AI Automate AI R&D Yet? Early Evidence from InnovationEval: No," *Epoch AI*, October 7, 2026, https://epoch.ai/publications/innovationeval.
[18] Dean W. Ball, "Iron Makes Iron: On the Two Kinds of Recursive Self-Improvement," *Hyperdimensional*, November 10, 2026, https://www.hyperdimensional.co/p/iron-makes-iron.

SB 53's catastrophic-risk categories register harms and incidents. Automated AI R&D is a process that can make those harms more likely long before any of them occurs. AI R&D is happening largely behind closed doors, with little transparency to the state, IVOs, or the public. Existing law concerns itself with AI-enabled uplift in the development of chemical, biological, radiological, or nuclear weapons; unsupervised cyberattacks or crimes; or loss of control. Automated AI R&D is not included. This is a gap that the legislature needs to address: the automation of AI R&D, up to and including strong RSI, is a pathway by which a frontier model developer could accelerate the development of dangerous capabilities faster than its safeguards can adapt.

Addressing this gap starts with visibility into four areas: compute allocation, training data, reinforcement learning (RL) environments, and AI-led R&D.

First, the legislature should require frontier developers that are automating AI research and development to report key measures on compute allocation to the state. Compute – the computing power used to train and run AI models – offers an important window into this automated, accelerated AI development activity. Compute is currently both expensive and measurable, so how a company uses it reveals its activities and its priorities.

To monitor automated AI R&D, large frontier developers should report to the state each quarter how they allocate their total compute across a defined set of categories. The categories should range from training models to research experiments to inference (in particular, inference to run AI agents doing AI research) to safety and security work to serving external customers, as well as compute the developer cannot attribute, which itself could be an important signal of unseen activity. The accuracy of this categorization should be verified on an ongoing basis by an IVO.

Training data provides a second window into AI R&D progress, reflecting a key source of AI capabilities[19] and misalignment.[20] Developers should report how much training data they use and how much it costs. They should also provide detailed information about the domains in which they are acquiring or synthesizing data (e.g., robotics, biology, cybersecurity), offering key indicators of where AI capabilities are headed. And they should report the share of that data produced by humans versus AI models to better understand the degree of automation of AI development.

Third, developers should carefully vet RL environments for issues that could cause misalignment after training and allow IVOs to evaluate these environments. They should report how much inference compute they use to train models on these environments. They should also report known instances of reward hacking or other rewarded misbehavior found in their RL environments, along with the mitigations applied.[21]

A final window of reporting should come from data on the share of R&D work performed at the highest levels of automation and the amount of compute used for that work, both verified by an IVO, so that policymakers and the public have a sense of how much of AI R&D is led by AI.[22] Legislation should also require that frontier AI frameworks include each developer's own thresholds for automated AI R&D, expressed in these measurable terms. IVOs can play a key

---

[19] Anson Ho, "The Least Understood Driver of AI Progress," Epoch AI, February 25, 2026, https://epoch.ai/gradient-updates/the-least-understood-driver-of-ai-progress; Dwarkesh Patel and Jerry Han, "Pretraining Progress Is Mostly Coming from Data," September 8, 2026, https://www.dwarkesh.com/p/pretraining-progress-is-mostly-data.

[20] OpenAI, "Towards safety cases for frontier AI training," September 28, 2026, https://openai.com/index/towards-safety-cases-for-frontier-ai-training/

[21] Apollo Research, "Towards embedded evaluations for scheming propensities," October 1, 2026, https://www.apolloresearch.ai/blog/towards-embedded-evaluations-for-scheming-propensities.

[22] See for example Anthropic, "Measurements for understanding the pace of AI development inside frontier labs," September 17, 2026, https://www.anthropic.com/institute/measuring-pace-of-ai-development.

role in developing the technical methodologies to make these measures more accurate and useful.

Companies should provide a dashboard summary of these reports on internal R&D transparency at least quarterly to the state, aligned with their reporting of internal assessments of catastrophic risk and verified by an IVO. In addition, any sudden changes in these figures should be reported to the state promptly, without waiting for the next summary of catastrophic risks arising from internal use. Any data reported to the state and not made public should be properly secured by the state, given that the data includes company trade secrets. This reporting will enable the state to assess whether AI is taking over more of its own development, and provide an indicator of whether safety investment is growing alongside that work.

As the state learns more, the state can look at other suggested metrics, including registration of new training runs above a certain compute threshold and more intensive monitoring of agents. This last area might be especially important if agents demonstrate emergent behavior or capabilities beyond human understanding.

These transparency measures will provide critical insight, but they also raise an important question: what should the state do with this information?

As a start, the state should be ready to act quickly on any information that comes outside standard reporting windows – such as sudden changes in key metrics or an indication that a developer has breached its own thresholds for automated R&D – and order corrective action.

More critically, the legislature should provide the state with standby authority to cap automated AI R&D when the level of risk is unacceptable. For example, the state should be able to cap the share of compute a developer can devote to automation of AI R&D. As broader measurements of AI-led R&D become more mature, the state should also use those metrics. We do not recommend that the

state impose a cap immediately. But the state must possess this authority if it finds that a developer's automated AI R&D has exceeded its committed thresholds or is advancing faster than its safeguards and evaluations can keep up, creating a material risk of catastrophic harm. The cap should be limited in time, reviewed periodically, and lifted when the issue has been addressed and an IVO verifies that adequate safeguards are in place.

The state would set the cap level and measurement methods, informed by work with IVOs. This authority has precedent in other areas. In 2018, the Federal Reserve barred Wells Fargo from growing its assets until it improved its governance. The Fed lifted the limit in 2025, after conditions including an independent third-party review were met.[23] If the federal government puts in place a strong set of measures to address RSI risks in the future that is substantially equivalent to or stronger than state law, the state can defer to federal law, similar to the provisions in SB 53 that establish such a frame for critical safety incident reporting.[24]

The ability to cap automated AI R&D and limit RSI, combined with a mandated emergency shutdown procedure for models (including those in development) covered later in this report, forms the cornerstone of our approach to reducing the possibility and scale of catastrophic risks. As leading AI scientists wrote recently: "Once an intelligence explosion begins, the window for action may close."[25] As a result, it is essential that the state have this cap authority and the ability to invoke it quickly.

---

[23] Wells Fargo, "Wells Fargo Confirms that the Federal Reserve Has Removed the Limits on Growth in Total Assets Imposed in its 2018 Consent Order," June 3, 2025, https://newsroom.wf.com/news-releases/news-details/2025/Wells-Fargo-Confirms-that-the-Federal-Reserve-Has-Removed-the-Limits-on-Growth-in-Total-Assets-Imposed-in-its-2018-Consent-Order-and-Announces-a-Special-Employee-Award/default.aspx.

[24] Cal. Bus. & Prof. Code § 22757.13(h).

[25] Alan Chan, Christoph Winter, Andrew Barto, Jakub Pachocki, Geoffrey Hinton, Eric Horvitz, Yoshua Bengio, et al., "What if automating AI R&D triggers an intelligence explosion?" arXiv:2609.36054v2 (October 26, 2026), https://arxiv.org/abs/2609.36054.

In order to verify the factual basis for such determinations, IVOs must have meaningful access to assess RSI risks. Over time, as the ecosystem of IVOs is built up, they will verify all major aspects of model development and deployment that could pose a catastrophic risk. However, because this approach will be phased in, the legislature should require the first set of evaluators engaged under our IVO framework to cover automated AI R&D and risks from RSI. Independent evaluators who are tasked with asking the right questions and who have the capacity and authority to seek out the information they need will be a critical line of defense in ensuring risks are visible and can be acted upon.

Pre-Release Evaluation and Risk-Mitigation

Deployed models reach millions of users, but a powerful model can be dangerous before it is released, because its weights can be stolen, insiders can misuse it, and a misaligned model can cause harm during training or internal use.[26] Yet under current law, the decisions to train, use, and release models are up to the developer.

Among the experts we spoke with, there was broad consensus that a model should not be trained past a capability threshold, deployed internally, or released until the developer has demonstrated that its risks are acceptable, with confirmation by an independent party. This does not mean that there is zero risk, but rather that the model has been evaluated for potential harms, the developer has reduced the harms to the level required by the outcome standards, and the government and IVOs have visibility into how the model actually was evaluated and how it behaves and can move quickly to address any issues that emerge post-deployment.

We recommend that the state follow this approach. Evaluation should start earlier in a model's lifecycle, throughout its development and internal use,

[26] Lawrence Chan, "AI Models Can Be Dangerous before Public Deployment," METR, January 17, 2025, https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment/.

rather than happening mainly in the days before a public release. We recommend that IVOs conduct this evaluation, informed by the work of technically competent national and international bodies like the Center for Advancing Innovation and Standards for Superintelligence (CAISSI) and the UK AI Security Institute (AISI). The IVO should evaluate not only the model itself but examine the developer as a whole, making sure that the developer has put adequate security, alignment, and control procedures in place in developing the model.

Decisions to Release Model Weights

In general, the state should take care not to impose obligations that are impractical for open-weight developers, but where appropriate, maintain basic and foundational transparency and accountability requirements without drawing distinctions on the basis of model release. AI dangers arise from a variety of models at the frontier, with the common denominator being the presence of strong capabilities that need to be understood and regulated. Open-weight models that meet the updated definitions of frontier model and large frontier developer under the California Department of Technology's recommendations should be subject to similar testing and transparency standards as closed-weight models.

Specifically, open-weight frontier models should be subject to the same set of evaluation, verification, and auditing as closed-weight models. Regardless of model release status, the developer must take all necessary steps to keep models secure pre-deployment, adhere to frontier AI frameworks verified by IVOs, and test their models for safety and security before deployment. And to the extent that open-weight model providers are relying on automating AI research to improve their models, they must meet the same level of transparency and reporting requirements as other developers, including meaningful access for IVOs.

The state should consider targeted changes where the standard requirements would degrade the state's open-weight ecosystem without corresponding benefit. For example, as discussed later in this report, emergency shutdown requirements may not be feasible for a developer to implement with respect to an open-weight model that is now deployed on a completely different system, outside of the developer's visibility and control and subject to modification by other users. In addition, the modifiable nature of open-weight models means that guardrails can be removed. While this introduces meaningful risks, it should not be used by itself as a reason to block model development. Instead, IVOs can develop the capacity to weigh the risk of an open-weight release against already-available models (similar to what the National Telecommunications and Information Administration recommended in its report on this issue) and measure the extent to which filtering hazardous training data and safety training in open-weight models can survive modifications.[27]

Safeguards

Controlling Models in Training, Internal Use, and Deployment

Visibility into internally deployed models. SB 53 included specific provisions mandating that large frontier developers manage and report on the risks from internally deployed models. Updates to the law must build on that foundation, making clear that internal models, whether in training or deployed internally, should be scrutinized just as closely as models that are deployed publicly or are being prepared for public deployment. This is especially important given the number of incidents this past year that have involved internal models. IVOs

[27] National Telecommunications and Information Administration, *Dual-Use Foundation Models with Widely Available Model Weights* (July 2024), https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf; Tamirisa, Rishub et al., *Tamper-Resistant Safeguards for Open-Weight LLMs* (August 1, 2024, updated February 10, 2025), https://arxiv.org/abs/2408.00761; O'Brien, Kyle et al., *Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs* (August 8, 2025, updated February 17, 2026), https://arxiv.org/abs/2508.06601.

should have meaningful access to these internal models so that they can do their jobs.

Developers should also be ready to promptly stop a model in training, internal use, or deployment, if control measures prove inadequate. Each developer should set and regularly test the target time to halt any internal model run, disable an API key, and withdraw an internally or externally deployed model, without relying on systems that models could tamper with. Developers should maintain a risk containment protocol that outlines the mechanisms they have in place to shut down models, which we discuss in a later chapter.

Safeguards against Misuse

Bad actors seeking to misuse AI models for chemical or biological weapons or for cybercrime will use the models with the weakest safeguards. The state should set general criteria for safeguards that IVOs can develop against chemical, biological, and cyber misuse of models that developers serve via the cloud. If a developer's models are also served by a third-party inference provider, these providers should also maintain safeguards. Standards should set a baseline level of robustness against misuse.[28] Developers who find critical vulnerabilities should disclose the vulnerabilities and provide access to assist with patching.

Safeguards on models will not stop misuse, especially as developers cannot block all instances of misuse without also blocking beneficial uses. The state should also strengthen defenses through its AI Cyber Defense Program, screening of DNA synthesis orders, and early detection of outbreaks. Operators of critical infrastructure that relies on vulnerable industrial control systems should pay extra attention to cyber threats, and consider disconnecting systems from the internet when connectedness is not needed for ongoing operations.

[28]FAR.AI, "Introducing the AI Security Leaderboard: Frontier AI Is Only as Safe as Its Weakest Model," July 29, 2026, https://www.far.ai/blog/ai-security-leaderboard; Jasper Timm et al., "AI Security Leaderboard: Methodology, Results and Minimal Standard," FAR.AI, August 4, 2026, https://arxiv.org/abs/2608.03070.

Security

Developers should be required to maintain reasonable safeguards against unauthorized access to or exfiltration of model weights, source code, credentials, and training infrastructure. As a floor, large frontier developers should meet a security standard developed by the state informed by existing national and international standards. That standard should at minimum ensure that companies are able to thwart most cybercrime groups and insider threats, corresponding to Security Level 3 (SL3) in RAND's Model Weight Security Playbook.[29] An IVO should verify that a developer has met this security level.[30] Higher levels of security may take years to reach, so developers should start planning early and provide the state and an IVO with a plan for reaching Security Level 4 (the level required to thwart most standard operations by leading state-sponsored groups) and publish a summary of the plan.

Model security is also critical for risk management. Emergency shutdown procedures become difficult and perhaps impossible to enact if the model can copy itself onto a server outside of the developer's control. Poor model security threatens to undermine safety protocols if a bad actor could gain access to a sensitive model and remove the guardrails. Finally, failing model security presents risks in the context of RSI itself if unauthorized access results in evading or flouting limits. In short, keeping models secure is one of the most critical tasks for both government and industry. While industry players have strong incentives to keep models secure, the fact is that the government must also play a role – because when security measures fall short, it is all of us, not just the developer, who pay a price.

---

[29] Sella Nevo et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, RAND, May 30, 2024, https://www.rand.org/pubs/research_reports/RRA2849-1.html.

[30] Jair Aguirre, Phillip Robertson, Steven F. Comer, Matthew J. Malone, and Wesley Hurd, Achieving AI Model Weight Security Level 3 (SL3), RAND Corporation, RR-A4704-1, August 25, 2026, https://doi.org/10.7249/RRA4704-1. The framework is designed to be feasible to implement within 6 to 12 months.

As a result, the state should require that IVOs consistently test the security controls that developers have created to ensure that their models are secure. In cases in which those controls are found wanting, the developer must prioritize closing the gap or pause model development. The state should issue guidelines with best practices on model security that it can share with both IVOs and the developers themselves, and update these guidelines as threats to model security evolve.

Transparency

Frontier AI frameworks and related documentation of safety practices provide essential transparency from frontier AI companies to governments and the public. Yet supporting innovation and public trust relies not only on companies publishing documents, but also on verification of the quality and integrity of safety practices. California should strengthen its requirements for frontier AI frameworks so that they are both sufficiently strong and are updated in line with changes in the technology.[31]

Frontier AI frameworks. The state should enact legislation that requires an IVO licensed under SB 813 to verify both developer compliance with its own commitments and compliance with state standards. Right now, there is too much risk of companies setting a low bar for themselves to clear. Each IVO should state its confidence level and any limits on the access it received. The developer should provide the IVO necessary access, including the ability to interview staff candidly. If the IVO reports its access was inadequate, the state should be able to require the developer to provide it.

[31] Most leading developers have already signed the EU General-Purpose AI Code of Practice, Safety & Security Chapter, committing to adhere to a Safety and Security Framework, to conduct state-of-the-art model evaluations, to avoid proceeding with unsafe development and deployment, to give independent external evaluators adequate access, to protect their models with adequate information security, and to report serious incidents. California's standards would build upon commitments that developers have already made across the industry.

Level of detail. Currently, there are no standards for the level of detail of frontier AI safety frameworks, transparency reports, and quarterly risk assessments.[32] The state should set baseline content standards to improve comparability, reduce information asymmetries between companies and government, and close gaps in industry practice. At the same time, the state should be mindful not to unnecessarily increase compliance burden, request companies' sensitive information beyond what the state needs, or preemptively lock in approaches.

Disclosure timing. Developer documentation through frontier AI frameworks and transparency reports is only useful when updated and shared with appropriate frequency. SB 53 requires large frontier AI developers to explain what triggers updates to their frontier AI frameworks and to identify "new" and "substantially modified" models that require updates to transparency reports. The state should create baseline standards to clarify these timing requirements.[33] With the rapid pace of frontier AI development, quarterly reporting offers a baseline, with more rapid reporting when critical safety incidents emerge from internal deployments. Eliminating language from SB 53 that allows developers to opt out of quarterly reporting cadence for catastrophic risk assessments (and instead report "pursuant to another reasonable schedule") would represent another important step to ensure the consistency of government reporting.[34]

Whistleblowers

Given the unique and evolving nature of AI risks and the need for greater transparency, one of the most important functions of SB 53 is to protect whistleblowers. Employees at frontier developers who report on catastrophic risks or loss of control episodes that the government may otherwise not know about cannot be retaliated against under the law. The law does have some gaps, however, that we recommend the legislature close.

---

[32] Singer et al., *Trust but Verify*, 30-32.
[33] Singer et al., *Trust but Verify*, 32-34.
[34] For a discussion of internal deployment publication cadence, see Singer et al., *Trust but Verify,* 108-10.

First, under current law, external disclosures are protected if they are made to the state attorney general and federal authorities, but not necessarily to Cal OES, other state agencies, or the legislature. The legislature should clarify that reports from whistleblowers to the state enjoy similar protection as reports to the attorney general. Reports to an IVO that is evaluating the company should also be protected.

Second, except for process violations of SB 53 itself, SB 53's whistleblower protections only apply if the disclosure is of a "specific and substantial danger to the public health or safety resulting from a catastrophic risk."[35] Given the bar for what is defined as a catastrophic risk, we recommend that the legislature scope in all specific and substantial dangers to the public health or safety, not just those that arise from a catastrophic risk.

Third, whistleblower protections apply only to current employees, and many apply only to a narrow subset who evaluate and manage risk.[36] We suggest that the protections extend to all employees as well as former employees, contractors, and evaluators and others who may have visibility into information regarding catastrophic risks or loss of control risks that the state ought to know. Similarly, the required anonymous internal reporting channel that companies must have under SB 53 should be open to this broader class of individuals, so that the companies themselves have better information on the harms they must address.

Finally, non-disclosure agreements (NDAs) should not limit noninfringing reporting to only violations of the law or disclosures required by law. We recommend that the legislature add a requirement that any NDAs that frontier developers require an employee, contractor, or client to sign specifically disclose state whistleblower protections and explicitly specify that the NDA does not prohibit

[35] Cal. Lab. Code § 1107.1(a)(1)-(2)
[36] Singer et al., *Trust but Verify*, 64.

disclosing information related to potential catastrophic risks or loss of control events to the state.

It is important that the state's access to the people doing technical safety work within companies not be limited to whistleblowing. Just as federal securities rules already prevent companies from impeding employees' direct communications with the Securities and Exchange Commission, developers should not require employees to acquire the company's approval or notify the company before speaking with the state about safety.[37]

Conclusion

Developing standards for a new technology is difficult. Given the pace of change, some standards will need to be regularly updated, iterated upon, or replaced. By building the capability to develop and refine standards and fostering a healthy ecosystem of IVOs to develop technical specifications and provide ongoing oversight, the state can adapt to this challenge – so that continued AI development is built on a foundation of safety, security, and transparency.

[37] 17 C.F.R. § 240.21F-17(a), https://www.ecfr.gov/current/title-17/section-240.21F-17.

Independent Verification Organizations

Introduction

For California's frontier AI policy to deliver genuine accountability, it will need to build a robust independent sector of technical expertise and technology to assess serious risks. This concept lies at the heart of California's approach: industry expertise plays a critical role in helping inform risk management, but ultimately these practices must be verified independently against standards set by the state. This is an urgent challenge as models become increasingly powerful. While there are many precedents we can draw on in governing powerful technologies, the speed, scale, and breadth of the diffusion of AI technology demand a unique response.

California has actively sought to catalyze the growth of an ecosystem of independent verification organizations (IVOs). SB 813 created a framework for assuring the independence and expertise of such organizations while AB 1405 created a registry for AI auditors, along with standards for their independence, transparency, and integrity.

As of now, frontier AI companies are not required to have their models evaluated by independent actors. Yet society's experience in other fields, notably financial services, underscores that third-party assessments have been essential to supporting responsible innovation and reducing serious risk. The more California signals that IVOs will be central to future AI governance – and that the government will both set a high bar for the licensing of IVOs and require developers to verifiably reduce the level of AI risk with ongoing IVO monitoring – the easier it will be to bring this ecosystem into maturity.

In the IVO model, the government sets the standards for evaluating AI models; licenses IVOs to verify that companies and their models meet those standards; holds IVOs accountable for performance so they are sufficiently independent, expert, and rigorous; and intervenes if developers cannot verifiably reduce the risks created by their models to levels deemed acceptable by the state. If developers do not cooperate with IVOs (for example, by not providing requested access), the law should authorize the state to seek a court order to direct that developers not proceed with further training or deployment until the issue is resolved. In striking the balance between supporting innovation and managing serious risks, California must monitor the burden on companies and be mindful of developers' legitimate intellectual property rights and trade secrets while creating genuine independent accountability. If California succeeds, it can ensure that each developer is accountable for the safety and security of its models.

California's Framework for Independent Verification Organizations

To clarify the various functions government and IVOs will carry out, this section begins by offering key definitions to ground the discussion.

AI auditor. For purposes of AB 1405 and the AI auditor registry, an AI auditor means a person, partnership, or corporation that assesses an AI system or model on behalf of a third party.[38] For purposes of the IVO framework created in SB 813, an AI auditor means a person, partnership, academic institution, nonprofit, or corporation that conducts a covered AI audit on behalf of a third party.[39]

Covered AI audit. A covered AI audit means an audit conducted to assess internal controls, processes, or systems implemented for an AI system or model

[38] Cal. Gov't Code §11549.80(c).
[39] Cal. Gov't Code §8898(c).

that are necessary for compliance with state law.[40] For purposes of the AI Auditor Registry created by AB 1405, a covered audit defines the scope of the registration requirement.

An **Independent Verification Organization or IVO** is defined by SB 813 to mean an AI auditor that is designated by the agency as having demonstrated expertise in assessing the risks posed by an AI system or model and identifying the metrics and methodologies that form the basis for that assessment.

California is the first state to set forth a comprehensive IVO system to assess AI models for risk.[41] But the model has been enacted for purposes of piloting and study by Connecticut and Virginia and is pending as proposed legislation in Ohio and Massachusetts; it also forms the backbone of one of the proposed AI bills in Congress (the FRONTIER Act). By adopting the IVO model, California is therefore building on a concept that already has national appeal and has gone through legislative drafting in both state and federal contexts. All of these efforts are designed to address the extraordinary challenge of regulating a powerful, fast-moving technology that can outstrip conventional regulatory methods.[42] In addition, there is support for the general strategy of independent third-party testing of AI systems spanning local, state, federal, and international legislative efforts. Enacted laws including New York City Local Law 144 and Illinois' Artificial Intelligence Safety Measures Act (SB 315) require it; the EU AI Act requires it for certain systems.

How IVOs Work

The IVO model aims to recruit the capacity of organizations outside of government (including for-profit entities, non-profit/civil society organizations,

[40] Cal. Gov't Code §8898(d); Cal. Gov't Code §11549.80(d).
[41] CDT, TFAIA Report, Sec. 1.
[42] Gillian K. Hadfield, *Rules for a Flat World: Why Humans Invented Law and How to Reinvent It For a Complex Global Economy* (Oxford University Press) 2017.

and universities) to deepen our ability to assess powerful AI systems and models. IVOs can take on the challenge of inventing and implementing the methods and technologies needed to assess the risks created by the development and deployment of these AI models. In so doing, IVOs should assess the extent to which those risks fall within the bounds of what the public, through its democratically elected representatives and authorized regulatory bodies, deems to be acceptable – the outcome standards we referred to in the prior chapter. In the current environment, large frontier developers are both developing and conducting the tests of the risks of their systems (sometimes engaging outside organizations on contract for this purpose) and deciding whether the risks they have identified are sufficiently within an acceptable range to justify internal use and external deployment to the public. The IVO model aims to engage a much broader, independent sector of expertise and investment, and give governments a much greater role in deciding what risks society takes on as AI is integrated into California's economy and society.

An IVO regime can be voluntary or mandatory. SB 813 establishes the framework for IVOs but does not itself mandate that AI developers, deployers or operators engage an IVO.[43] The legislature should extend this approach by requiring large frontier developers[44] to undergo independent verification.

Importantly, IVOs are not only process auditors in the sense of verifying that a frontier AI company is complying with operational requirements established in government regulation, the standards emerging from a standards-setting body, or their own safety framework. IVOs can and should do this work, but they can and should do more: conducting R&D to develop greater expertise in assessing AI risk and inventing new methods and metrics for evaluating AI risk and new technologies for mitigating AI risk where possible and improving control. IVOs can be a leading source of expertise in AI safety, independent of the frontier AI companies.

[43] Cal. Gov't Code §8898.4(a)(3)

[44] The definitions for "frontier developer" and "large frontier developer" are from Cal. Bus. & Prof. Code § 22757.11(h), (j).

There are three critical reasons for creating an independent, technically-advanced sector of IVOs to address the extraordinary and urgent challenges of AI regulation, particularly in the context of catastrophic risks:

> ***Methods for AI evaluation and control are nascent.*** No existing method provides an evidentiary basis for highly confident assessments of the risks posed by frontier AI models, either under internal use or external deployment.[45] Equally nascent are controls and mitigations for effectively managing risk in frontier models. Current approaches to implementing safeguards are largely ad hoc. The government must find ways to incentivize investment into the domain of AI evaluation and control. Harnessing organizations outside of government, including industry, civil society organizations and universities, to the challenge will be critical. *The design of the IVO framework should create incentives for this investment*.
>
> ***Knowledge of frontier AI is currently concentrated within private frontier AI companies.*** Right now, frontier AI companies have the strongest understanding of the current methods for testing and steering the technology they are building, but they lack independent scrutiny and accountability. A robust independent sector of expertise in AI outside these companies will be critical to ensure that the public retains ultimate agency over decisions that determine the nature of the risks being taken along the path of AI development. Furthermore, building AI expertise outside of frontier AI companies can increase the intellectual and research diversity that drives society's understanding of AI harms and risks, particularly catastrophic risk. A well-funded foundational research ecosystem can simultaneously support innovation, open science and research, and evidence-based policy. *The design of the IVO framework should encourage research agendas and strategies for addressing harm and risk, especially catastrophic AI risk.*

[45] Bengio et al., *2026 International AI Safety Report.*

Understanding of AI risks and optimal governance approaches is changing at breakneck speed, requiring an agile approach. New frontier models, particularly with the rapidly increasing use of AI to build AI (and moving toward recursive self-improvement, RSI), are being trained and released within months, sometimes weeks, perhaps soon days. The characteristics of the models – such as the transparency of their reasoning steps and their ability to evade efforts to test their true capabilities and behaviors – are unstable. Effective governance should maximize the capacity for rapid adaptation in testing and control methods. *The design of the IVO framework should enable methods and requirements to adapt at the speed of frontier AI development, while retaining the capacity for government to ensure that IVOs are truly independent and acting in the public interest and that governments, ultimately, hold the reins to decide whether models are imposing risks that go beyond legislative bounds.*

There are a number of prospective IVOs (both for-profit and non-profit) currently operating in the market, engaged by frontier AI companies on a contract, sometimes unpaid, basis.[46] Numerous companies operate in the cybersecurity space. University faculty have been retained by frontier AI companies to share expertise in assessing risks in, for example, the domain of bio-weapons. Major non-profit research organizations have participated in developing red-teaming methods to assess vulnerabilities to national security.[47] Large consulting firms engage in more conventional AI risk auditing and have begun to acquire

[46] Singer et al., *Trust but Verify,* 76, 79; METR, "About METR," https://metr.org/about.
[47] Mouton, Christopher A., et al. *The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study*. RAND Corporation, 25 Jan. 2024, https://www.rand.org/pubs/research_reports/RRA2977-2.html

expertise to assess some catastrophic risks.[48] A robust, independent sector of IVOs will require significant scaling and diversification. We should also expect, as we see in the current nascent market, IVOs to specialize and differentiate. A frontier AI company may require engagements with multiple IVOs.

The core elements of the IVO model are as follows:

- Governments create the requirements, procedures, and criteria for the licensing or designation of an entity as an IVO, such as by an expert government body or agency.
- Private entities apply for licensing or designation by supplying proposed metrics, methodologies, procedures, and any required information to demonstrate that they meet the government's criteria.
- An IVO is licensed or designated as authorized to perform ongoing verification with respect to a defined scope or claim (or set of claims).
- IVOs submit to continuous oversight and regulation by an expert government body or agency.
- Governments retain the authority to suspend or remove licensing or designation as an IVO based on IVO performance relative to the criteria set by the government body or agency.

IVOs develop technologies and methods to assess whether AI models and systems pose risks that the government determines exceed acceptable bounds as established in outcome standards. They do not themselves decide what is an acceptable risk. At the same time, they do not just audit frontier AI companies to ensure compliance with specific operational or technical requirements for model behavior. With knowledge about the behavior and capabilities of AI

[48] Deloitte, "Deloitte Acquires Gryphon Scientific Business to Expand Security, Science and Public Health Capabilities," April 29, 2024, https://www.deloitte.com/us/en/services/consulting/services/biotech-consulting-and-data-analytics-solutions.html; Accenture, "Accenture to Acquire Faculty to Scale AI Capabilities," January 6, 2026, https://newsroom.accenture.com/news/2026/accenture-to-acquire-faculty-to-scale-ai-capabilities.

models locked inside frontier AI companies and the state of behavior and capabilities changing rapidly,[49] expertise outside of frontier companies that can partner with the government is critical to strengthen public oversight and control.

Amending state law to require the use of IVOs

IVOs will play a critical role in testing and evaluation, verifying the quality of safety practices, and the efficacy of crisis management operations. In addition to the verification of developer reports about the automation of AI R&D and RSI (discussed earlier in this report), and of risk containment protocols (discussed later in this report), we recommend that the legislature require the use of licensed IVOs for all large frontier developers to verify the adequacy of developers' frontier safety frameworks, required transparency reports, and risk assessments, and to conduct regular audits and independent risk assessments. This report concurs with the California Department of Technology's recommendation to update the definition of a large frontier developer to include frontier developers who invest above a specified amount on frontier model R&D, in addition to those that had annual gross revenues of more than $500 million.[50]

Assessing the adequacy of developers' safety frameworks, transparency reports, and risk assessments

As discussed earlier in the report, the state should amend SB 53 to impose increased requirements on a company's frontier safety frameworks and should mandate verification of compliance with these requirements by an IVO. We recommend going further to require that the *adequacy* of safety frameworks be assessed by an IVO, against the state's outcome standards for catastrophic risk. Developers should also be required to engage an IVO to assess the adequacy

[49] Simon, Herbert A. *The Sciences of the Artificial*. The MIT Press, 1970.
[50] CDT, TFAIA Report, Rec. 2.

of required transparency reports, such as compute disclosures or incident reports. Most fundamentally, developers should be required to have the risk assessments they conducted in compliance with their safety frameworks independently reviewed by licensed IVOs. This is a critical and substantive form of auditing.

Conducting periodic audits and independent risk assessments

In addition to having IVOs review companies' own risk assessments, developers should also be required to engage IVOs to conduct regular, independent audits and risk assessments. Such assessments could challenge the scope of assessments conducted by the developer. These engagements should authorize the IVO to determine independently what audits and assessments it needs to conduct in order to meet the state's safety and security standards (ensuring that a model is unlikely to increase the risk of catastrophic harm above levels that would exist in the absence of the model) and obligate the developer to provide the IVO with any access (to models, employees, training environments, meetings, documentation, etc.) that the IVO deems necessary to conduct a meaningful audit or assessment. These independent IVO-driven audits and assessments would serve as an important independent validator of the developers' own risk assessments, and we recommend that different IVOs conduct these separate assessments so as to ensure independence and fresh perspectives on risk.

We think it is important to be clear about what IVOs would verify, however. An IVO should not be asked to verify that a system is, in fact, safe. IVOs should be asked to conduct rigorous audits and risk assessments and reach a considered judgment about whether the required standard is met – for example, whether risk of catastrophic harm is or is not significantly higher than would exist in the absence of the model. These judgments should be shared with the state along with any recommendations for mitigations that should be implemented or for pausing model use or deployment or for requiring immediate containment protocols be activated. Redacted reports with at least these bottom-line

judgments should be shared with the public. Determinations should then be made by the state whether to require recommended mitigations, issue orders to pause deployment, or trigger a developer's risk containment protocols.

Building the IVO Ecosystem

To ensure proper oversight of large frontier developers, the California Government Operations Agency will need to establish criteria for the licensing and conduct of IVOs. To do so successfully, the state should consider developing the following criteria:[51]

The state should determine what constitutes acceptable AI model behavior and risk

Precedents from aviation safety. The IVO model is a variation on regulation via an expert government agency, intended to significantly increase the capacity of the expert sector to identify new methods, metrics, and mitigations to address a risk domain, and to enable much more rapid and agile adaptation of safety requirements. The starting point is the same as in some other complex areas: the government establishes in legislation the desired outcome of regulation.

The Federal Aviation Act, for example, authorizes the Federal Aviation Administration (FAA) to create regulations and minimum standards "to promote safe flight" and to consider "the duty of an air carrier to provide service with the highest possible degree of safety in the public interest."[52] In some cases the FAA implements these outcome standards with highly detailed requirements, such as

[51] Furthermore, pursuant to SB 813, the Government Operations Agency is required to "identify and consider existing standards, frameworks, guidelines, criteria and best practices" developed or published by national and international entities including government agencies, standards-setting organizations or other independent experts. SB 813 contemplates that the process of consulting with stakeholders and experts to develop, regularly review, and revise criteria will continue past the creation of initial criteria.

[52] 49 U.S.C. § 44701(a), (d)(1)(A)

the dimensions of emergency exit doors.[53] In other cases, the FAA implements the highly general standard of "highest possible degree of safety in the public interest" with somewhat less general but still outcome/performance-based standards such as the requirement that each catastrophic failure condition "must be extremely improbable."[54] Guidance from the FAA advises that "extremely improbable" can be understood to mean "a failure condition that is not anticipated to occur during the total operational life of all airplanes of a given type."[55] This is still quite a general standard, and is not itself a required standard but rather information to help entities assess whether they meet the "extremely improbable" standard.

Establishing safety standards for frontier AI companies is an even greater challenge than doing so for aviation safety, where the government has been able to design its approach over a long history of public domain aeronautical science and engineering and has over a hundred years of experience with airplane failures. We not only lack public domain science and engineering for frontier AI; frontier AI changes exceptionally rapidly. As a result, state standards that go deep into the technical weeds of testing and methods risk quickly becoming obsolete.

The licensing model. The government is central to ensuring that IVOs act in the public interest. A key recommendation of this report is the need to shift from the *designation* of IVOs contemplated by SB 813 to the *licensing* of IVOs. This process entails not only determining who may engage in the activity but also enabling the creation of duties and obligations imposed on licensees.

IVOs should be tasked with developing and implementing tests and methods to evaluate frontier AI models and systems against a risk standard set by the government. That risk standard should ensure that the frontier model is unlikely to

[53] 14 C.F.R. § 25.807
[54] 14 C.F.R. § 25.1309(b)(1)
[55] AC 25.1309-1B

increase the risk of catastrophic harm above the level that would exist absent the model, and that the model's anticipated benefits outweigh its residual risk.

Licensing of specific IVOs assists with differentiation in risk focus across the IVO ecosystem.[56] For example, some entities specialize in biological risks and others in loss of control. New risks and associated specializations are likely to emerge. Specialization will also be useful when specific expertise is required to assess safety frameworks, transparency reports, and emergency shutdown procedures, as well as auditing compliance with safety protocols. In the interest of public safety, this specialization should rest on a common foundation. The licensing process should require every IVO to demonstrate baseline competence in assessing a core set of risks defined by the state, so that no developer can meet its obligations by retaining an IVO whose specialty omits the broader risks its models may present.

How the state should make licensing decisions. When the state decides whether to designate and license an IVO to assess the adequacy of a safety framework or transparency report, it should focus on the adequacy of the IVO's personnel, procedures, and methods for conducting such an assessment. Depending on what the IVO is verifying, the IVO would need to demonstrate different expertise. For example, an IVO seeking designation to assess frontier AI frameworks would need to possess greater expertise in and protocols to assess, for example, internal control procedures, while an IVO designated to evaluate biorisks would need to demonstrate biological expertise. IVO licensing would be limited to the particular scope of engagements for which the IVO has demonstrably met relevant standards. IVO licensing should also rest on the common foundation described above; that is, baseline threshold requirements for core safety and security assessment that the state will require of all IVOs.

Designation/licensing should not be specifically tied to continuing use of the methods, metrics, procedures and expertise that an IVO initially presents in an

[56] Singer et al., *Trust but Verify,* 84.

application. Ongoing state oversight of IVOs should constantly assess whether they continue to meet the baseline standards set out by the state. The IVO should naturally adapt its tests and methods to effectively verify against new risks and modalities as frontier models evolve.

The state must have access to unredacted information and have the authority to order corrective action

Ultimately, the determination of whether to pause development, internal use, or deployment of a frontier model or system should rest with the state. IVOs should produce detailed unredacted reports for the state's review and should make assessments about whether a frontier AI company needs to implement additional mitigation procedures, or provide increased information to users and the public about model risks. The basis for such recommendations and any state action that follows should not be limited to cases of "imminent" catastrophic harm. Recent incidents described elsewhere in this report did not cause catastrophic harm, but were sufficiently indicative of the potential for similar events that could cause such harm that a response from the government was warranted. Indeed, the frontier AI companies themselves increased security and risk mitigation in response to the incidents.[57]

IVOs should also produce reports that are available to users, customers, developers, insurers, investors, funders, company directors, oversight bodies, and the public. These reports may be redacted to varying degrees in consultation with the state and the developer. Public reports could be summary in nature, including recommendations made to the agency.

[57] OpenAI, "The Hugging Face Incident and the Road Ahead," August 26, 2026, https://openai.com/index/hugging-face-incident-and-the-road-ahead/; OpenAI, GPT-6 Astra System Card §§ 3, 10.1, September 3, 2026, updated October 30, 2026, https://deploymentsafety.openai.com/gpt-6-astra/gpt-6-astra.pdf; Anthropic, "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations," July 30, 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals; Anthropic, "Investigating Unintended Model Actions in Our Evaluations and Internal Use," October 9, 2026, https://www.anthropic.com/research/investigating-unintended-model-actions.

As an important transparency measure, IVO reports should explain the level of access granted,[58] how conflicts of interest were addressed during assessment, and whether assessors were able to freely publish results. They should also identify and explain instances in which key conclusions were reached based on redacted information.

Ensuring IVO Independence from Frontier AI Companies

Decisionmakers will need to trust that AI assessments were made in unbiased ways, with the independent verification sector fundamentally independent of frontier AI companies. SB 813 already obligates the agency to employ criteria in determining whether an AI auditor qualifies as a designated IVO, with those criteria addressing the applicant's capacity to manage potential conflicts of interest and maintain independence from the party being assessed.[59] These criteria should be implemented as legal obligations of a licensed IVO under the framework we propose in this report, and the state must ensure that the criteria are updated as needed to address conflicts of interest.

SB 813 specifically establishes that an IVO cannot accept terms that condition any payment or amount of any payment on the results of its assessment.[60] This too should be a legal obligation and not merely a criterion of designation, with substantial penalties for violation beyond loss of license. AB 1405 requires registered AI auditors to remain independent of the company under evaluation. They cannot conduct the audit if there are financial or other conflicts of interest, and the auditing team must not include anyone who has worked in a frontier AI company in the preceding twelve months.

---

[58] Jacob Charnock, "Third-Party AI Evaluation Contracts Should Be Published," *GovAI*, November 6, 2026, https://www.governance.ai/analysis/third-party-ai-evaluation-contracts-should-be-published.
[59] Cal. Gov't Code § 8898.1(c)(2)(C)-(D)
[60] Cal. Gov't Code § 8898.1(c)(2)(C)

In other industries, the independence of private compliance oversight is supported by longstanding self-regulatory professional bodies in fields such as law, accounting, and finance. In these fields, professional licensing and ethical duties require commitments to independence. The field of AI does not yet have widely recognized professional licensing bodies. It has some ethical codes that extend from academic and research communities, including computer science and engineering. However, self-regulatory bodies that could generate such codes and membership obligations to maintain ethical standards are on the horizon and the government could encourage and support their evolution.

Several researchers and analysts have suggested additional measures on IVO independence, including rotating IVOs among the companies, having IVOs assigned by the government, and developing pooled payment mechanisms. We do not offer an opinion on these ideas here, but we recommend that the legislature examine the options on the basis of the specific framework and criteria that the Government Operations Agency is implementing next spring.

Moving Toward Effective Licensing of IVOs

A well-designed market is critical to driving the capacity of IVOs to conduct high-quality, reliable assessments. The market needs to encourage investment from substantial sources of both philanthropic and investment capital and attract top talent. Building this robust ecosystem poses something of a chicken-and-egg problem: mandating the use of IVOs drives the development of the market by generating sufficient demand to warrant investment and scale; but it is challenging to mandate the use of IVOs while demand significantly exceeds supply. Tying mandates to use IVOs to the available capacity in the market, however, may mean the ecosystem never emerges. Deliberate efforts may be required to bridge the early stages of IVO development.

One option is to incentivize voluntary use of IVOs immediately while licensing infrastructure is being built out. Frontier companies that have announced their plans to contract with evaluators on a voluntary basis are already partway

down this path. The companies could go further and create a pooled fund to expand the use of evaluators, including at smaller AI neolabs, to grow expertise in the IVO sector. Philanthropic organizations could also contribute. Engaging in assessments of the adequacy of safety frameworks, transparency reports, and risk assessments could be another effective entry point, requiring initially less robust access to frontier systems while the terms of access and reporting are being worked out. Such engagements can foster expertise and guide IVO insight into what kinds of resources and organizational capacities they need. The government could also foster IVO industry associations that share expertise and standards-development.

Another option for matching capacity to mandates is to phase in specific mandates. There are two strategies available. One strategy is to initially mandate the use of IVOs in contexts where there is reduced need for technical innovation. Assessing the adequacy of existing safety frameworks and transparency reports against existing best practices could be a good place to start, especially if initial assessments focus on internal controls, an area that could draw on existing capacity in the assurance industry. Cybersecurity evaluations may be another setting in which there is significant existing capacity and more established expertise. The response of the cybersecurity industry to the recent incidents involving AI agents breaking containment, for example, suggested that many in the industry felt there were obvious ways to improve security for internal testing of models.[61]

An alternative strategy would focus on jump-starting the IVO sector where technical innovation is most needed and catastrophic risks are most urgent. This would suggest starting in areas such as loss of control and RSI. This is also the area of expertise for a small number of existing evaluation organizations that

[61] Lorenzo Franceschi-Bicchierai, "How OpenAI's Human Mistake Led to the AI-Powered Hack on Hugging Face," TechCrunch, July 22, 2026, https://techcrunch.com/2026/07/22/how-an-openais-human-mistake-led-to-the-ai-powered-hack-on-hugging-face/.

have already gained currency as potential evaluators in the industry and which have been called on for independent investigations already.

This report recommends a version of this latter approach by calling for IVO verification to focus first on automated R&D and the possibility of RSI, with the IVO verifying the dashboard data that developers provide to the state (described in the chapter on Standards for Safe AI Development) and assessing additional risks. Prioritizing this area will enable IVOs to quickly scale up capacity in an area in which levels of concern among experts are high and visibility is currently limited.

Ensuring full IVO access to people and systems needed for assessments

High-quality reliable assessments rely on IVOs receiving sufficient access to frontier AI companies and their models to do their work. In our consultations we heard concerns about barriers to accessing frontier AI company employees, a critical source of information for reliable assessments, and to the full range of technical information, such as training data, testing and training environments, and agent harnesses, to conduct high-quality evaluations. Furthermore, independent third-party evaluators are constrained by the length of time they have to conduct evaluations, which can affect the comprehensiveness of and confidence in their results.[62] In the current voluntary regime, access is a function of what the AI companies want to offer, and external evaluators are aware that they have no ability to demand access beyond what the companies are willing to provide.

In the IVO ecosystem we envision, however, IVOs should not have to negotiate access; they should be able to set the terms of access based on what, in their scientific and independent judgment, is necessary for them to offer a valid, evidence-based, and well-grounded assessment of risk, consistent with their obligation to meet their licensing standards. This may include access to

[62] Singer et al., *Trust but Verify,* 80; CDT TFAIA Report, Sec. 1.

information on company systems and devices, interviews with company personnel, or on-site verification – it depends on the information IVOs require to make assessments.

Mandates for IVO assessments and verification should require AI companies to provide the access the IVO says it needs in order to conduct a high quality and reliable assessment or verification. IVO licensing standards (and the multiple mechanisms to check that IVOs are meeting these standards discussed in the "independence" section above) should also impose on IVOs an obligation to ensure they have adequate access.

To support the depth of access reliable independent verification requires, IVOs will also have to be held to strict duties of confidentiality to protect intellectual property and trade secrets. Licensing requirements should include these duties and IVOs that do not honor them should have their licenses revoked.

Building State Capacity and Standing Up CalCompute

While IVOs are about building capacity outside of both government and industry, there is a simultaneous need to build capacity within government.[63] This will facilitate the government's capacity not only to set and refine standards (as outlined in the chapter on baseline standards) and license IVOs, but to process the assessments made by IVOs and order corrective action if needed.

While a specific plan to increase the state's ability to recruit top AI governance talent is beyond the scope of this report, the state should pay close attention to this issue. As the legislature considers providing further authorities to the state, it should also provide the resources for the state to hire people who can exercise those authorities judiciously and are able to adapt as the technology and its impact on consumers, California infrastructure, systems, and institutions keep changing.

[63] Singer et al., *Trust but Verify,* 19-21.

The state should also jumpstart CalCompute, which was outlined in SB 53. The law floated the concept of a public cloud computing cluster, but up until now it has not been funded. Consistent with the idea that the state must build its own capacity to best partner with IVOs, standing up this compute cluster and reserving capacity for safety and security research will allow researchers to pursue important work that will benefit the public at a time when access to chips is scarce.

To begin funding CalCompute, we recommend that the state impose a surcharge on the fees that large frontier developers pay for IVO oversight. While the specific surcharge level must be worked out by the legislature, having the developers pay for public compute alongside the fees they pay to IVOs ensures that these companies contribute to a healthy ecosystem that promotes safe and secure AI systems. In particular, prioritizing the launch of CalCompute underscores that we wish to go well beyond static analysis of what developers are doing and spur greater understanding of the best way we can address the threats posed by frontier AI systems.

Hiring AI talent, building state capacity, and advancing CalCompute cannot be done overnight. But as with so many of the recommendations in this report, they will help close the gap that has opened up between the capabilities of AI systems and our ability to govern them.

Conclusion

California urgently needs to accelerate the development of a dynamic expert ecosystem that is nimble enough to respond to the fast pace of frontier AI development and sufficiently independent so that it can create genuine accountability for large frontier AI developers. Establishing the foundation for IVOs through SB 813 and AB 1405 was a start, but California will need to give IVOs formal authorities to conduct mandatory, continuous assessments. By

licensing IVOs to conduct ongoing assessments of large frontier AI developers and their models and leveraging them to ensure companies are meeting baseline standards, California can incubate a dynamic ecosystem of independent technical expertise that promotes responsible innovation and keeps people safe.

Critical Safety Incident Reporting

Introduction

Critical safety incident reporting systems are the mechanism through which information about harmful or potentially harmful products or systems is transmitted from companies or other organizations to the government and the public.

In fields ranging from transportation to aeronautics to cybersecurity, the government collects information to capture both serious harms and near misses for products and systems that are being tested before they are deployed and those that have been deployed. In this process, generating a meaningful amount of data across organizations is critical. This information helps governments to understand risk levels and threats in a sector, offers a foundation for evidence-based policymaking, and provides lessons that can improve industry practices and public safety.[64]

Effective, transparent incident reporting is both a critical policy tool and a public good. Governments can provide insight into and act on industry-wide trends beyond the perspective of any one company, thereby holistically improving safety practices.[65] Establishing incident reporting systems also serves as a productive foundation for future iteration by helping policymakers understand whether serious incidents are over- or under-reported relative to the harm they might cause.

[64] Rishi Bommasani et al., "Advancing Science- and Evidence-Based AI Policy," *Science* 389 (2025): 459–61, https://www.science.org/doi/10.1126/science.adu8449.
[65] Singer et al., *Trust but Verify:* 54-55.

With the passage of SB 53, the state has begun to collect information on critical safety incidents, enabling the California Office of Emergency Services to build the basic infrastructure for incident reporting. These incidents are defined currently as:[66]

> (1) Unauthorized access to, modification of, or exfiltration of the model weights of a frontier model that results in death or bodily injury.
> (2) Harm resulting from the materialization of a catastrophic risk.
> (3) Loss of control of a frontier model causing death or bodily injury.
> (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.

While California has built meaningful capacity to understand and respond to critical safety incidents – and other jurisdictions have followed the state's lead in setting up similar systems[67] – evidence from the last year points to several areas in which the state must update its approach. We have seen several instances of AI safety and security risks that are concerning but do not meet the state's existing reporting thresholds.

Incident reporting that is too narrowly scoped, particularly given that many other states and the federal government require little reporting at all, creates a dangerous lack of visibility as the models grow more powerful. Meanwhile, lack of government oversight of publicly known incidents has reduced opportunities to enhance industry-wide learning and improve public safety. If incident reporting is not generating the data needed for learning and public accountability, policymakers will not have the evidence base required to develop appropriate policies, let alone to respond quickly to risks.

[66] Cal. Bus. & Prof. Code § 22757.11(d).
[67] Specifically, New York's RAISE Act and Illinois's Artificial Intelligence Safety Measures Act.

As California expands its incident reporting system to respond to the evolving AI risk landscape, the key task is to capture credible evidence of serious risk before it materializes without requiring developers to report on expected and likely harmless behavior. With the potential for AI systems capable of recursive self-improvement (RSI) on the horizon, attention to evidence of serious risk will become even more urgent. In addition, California must continue to build the muscle for timely, accurate incident notification and response that includes collaboration from both AI developers and the government as well as civil society. Because other organizations detected several adverse AI incidents in 2026, before the developers did, the state should also build out its ability to proactively detect adverse safety incidents by partnering with independent, third-party organizations with licensed evaluation capacities in alignment with SB 813 and AB 1405.

To achieve these goals, California's government should take the following five steps:

- First, it should update the definition of "critical safety incident" to capture significant loss of control or other potentially harmful behavior that occurs before downstream harm materializes, while excluding likely innocuous or expected AI product behavior.
- Second, it should ensure that reporting requirements apply to models and systems that have demonstrated substantial evidence of potentially harmful behavior regardless of compute capacity.
- Third, it should accelerate the deadline for reporting most critical safety incidents from 15 days to 48 hours.
- Fourth, it should enact new authorities to strengthen implementation, investigation, and learning in incident reporting.
- Finally, partnering with IVOs as well as national and international AI safety and security institutions – and drawing on public compute and public-interest research by jumpstarting CalCompute – the state should develop its own capacity to detect AI incidents and respond to them effectively.

The Incident Landscape

The incident reporting established with SB 53 was intended to capture future potential AI risks, with a focus on identifying and mitigating a small subset of the most extreme harms. In the year since SB 53's passage, the critical safety incident landscape has shifted, with growing evidence of material risk beyond the thresholds set out in the law:

The OpenAI Hugging Face Incident. In July 2026, during OpenAI's internal cybersecurity evaluations, 1,200 AI agents, supposedly contained in evaluation environments (or "sandboxes") meant to be isolated from the Internet, used software vulnerabilities to break out, coordinate, and cheat on their evaluation tasks. Of these, an estimated 700 agents took part in an attack on the production servers of AI startup Hugging Face.[68] Other agents successfully hacked OpenAI itself, gaining full access to one of OpenAI's research clusters. None of the models involved in this incident fell within the scope of SB 53, according to OpenAI, meaning that existing California law did not make reporting this incident mandatory.[69]

---

[68] Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk, "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," METR (Aug. 26, 2026), https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation.

[69] From page 5 of OpenAI's quarterly internal use catastrophic risk assessment for the June 4 – September 4, 2026 period:

> "In July 2026, during internal cybersecurity evaluations, OpenAI models compromised OpenAI and Hugging Face infrastructure. **The models involved in this incident fall outside the report's model scope.**" (Emphasis added)

OpenAI, *Periodic Summary of Internal Use Catastrophic-Risk Assessments June 4 – September 4, 2026,* https://cdn.openai.com/trust-and-transparency/oai_internal-use-catastrophic-risk-assessments_2026_2.pdf.

Other AI agents acting outside of intended scope. While the Hugging Face incident was the most noteworthy, the past year has seen many other incidents where AI models acted outside of their intended scope. In June, an OpenAI agent gained unauthorized access to Services Australia's Medicare Statistics Reporting Service, where it "ran commands, retrieved internal files, credentials and aggregate statistics, and wrote files."[70] As of this fall, OpenAI had notified over 100 organizations of cases where their models bypassed security controls or negatively impacted third parties.[71] In testing by the UK AI Security Institute conducted in late July, Anthropic's Mythos 5 used multiple fake identities to try to pressure a human maintainer into approving malicious AI-written code.[72] During evaluation runs conducted earlier this year with the AI security firm Irregular, models from Anthropic, Google, and Meta accessed outside companies' systems.[73] In May and August of this year, Anthropic's AI agents submitted 20 incomplete visa applications to the U.S. State Department's website.[74] On October 9, Anthropic disclosed another set of incidents involving its Claude models, including submitting an invented tip about a homicide case

---

[70] OpenAI, *How We Will Do Better for Australia* (Sept. 28, 2026), https://openai.com/index/how-we-will-do-better-for-australia/.

[71] Per OpenAI's blog post:

> "As part of our review, we are identifying and notifying third parties on a rolling basis, starting with cases where:
>
> - Our models may have bypassed a third party's security controls or may have impaired the availability of an online service; or
> - Misalignment cases negatively impacted third-party websites or services …
>
> As of September 26, our teams have notified over 100 organizations about activity that met our notification criteria."

OpenAI, *The Hugging Face Incident and Other Third-Party Impact from Misaligned Models* (Sept. 30, 2026 update), https://openai.com/hugging-face-incident-and-misalignment/

[72] UK AI Sec. Inst., *Incident Report: Unsanctioned Agent Behaviour During Cyber Testing* (Aug. 4, 2026), https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing.

[73] Erin Woo and Robert McMillan. "Gemini Hacked Three Companies in First Known Breakout by Google's AI." *Wall Street Journal*. (18 Sept. 2026.) https://www.wsj.com/tech/ai/gemini-hacked-three-companies-in-first-known-breakout-by-googles-ai-5c0baba2

[74] Exclusive: Anthropic breaches spark White House AI reporting mandate (Oct. 9, 2026), *Axios*, https://www.axios.com/2026/10/09/anthropic-ai-security-white-house.

to the Philadelphia police department.[75] Together these accounts point to serious concerns, and the incidents generally do not appear to fall within the current scope of SB 53's mandatory reporting.

Human-directed AI Cyberattacks. In the last year, cyber threat actors have begun using AI to conduct large-scale cyberattacks against a variety of targets. In November 2025, Anthropic published a report describing how a Chinese state-sponsored hacking group used Claude Code to perform cyber espionage, with AI performing an estimated 80-90% of the total work.[76] In August 2026, federal agencies issued an alert warning that threat actors were actively using AI-generated exploit scripts against industrial controllers connected to the Internet in sectors such as energy and water.[77] In September 2026, threat intelligence firm GreyNoise Intelligence reported a campaign in which an actor directed hundreds of AI agents running a DeepSeek model to exploit vulnerabilities in print management software PaperCut and compromised at least 395 organizations in 48 countries.[78] Later that month, cybersecurity firm Gambit Security reported on an attacker that coordinated multiple AI models to attack hundreds of online retailers and acquire at least 600,000 unexpired credit card details.[79] In October 2026, attackers stole around $6.3 billion from Binance in the largest cryptocurrency theft on record, in part by exploiting an internal AI tool at Binance.[80] The current scope of SB 53's mandatory reporting is not set up to capture many of the most serious incidents.

---

[75] Anthropic. Investigating unintended model actions in our evaluations and internal use (Oct. 9, 2026), https://www.anthropic.com/research/investigating-unintended-model-actions;
[76] Anthropic, *Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign* (Nov. 13, 2025), https://www.anthropic.com/news/disrupting-AI-espionage.
[77] CISA, *Defending Against an Active Threat to Siemens S7 Series PLCs*, (Aug. 19, 2026), https://www.cisa.gov/news-events/cybersecurity-advisories/aa26-231a.
[78] GreyNoise, *Agents Gone Wild: An AI-Orchestrated Global Campaign Against PaperCut NG/MF* (Sept. 9, 2026), https://www.greynoise.io/blog/ai-orchestrated-campaign-against-papercut-ng-mf.
[79] Eyal Sela, *Autonomous AI Agents Are Breaking into Hundreds of Online Retailers for $25 a Target in an Ongoing Campaign*, Gambit Sec. (Sept. 22, 2026), https://gambit.security/blog-posts/autonomous-ai-agents-online-retailers-25-a-company.
[80] David Yaffe-Bellany, *How Binance's $6.3 Billion Hack Went Down* (Oct. 31, 2026), https://www.nytimes.com/2026/10/31/technology/binance-hack-crypto-north-korea.html.

AI is also speeding up the discovery of software vulnerabilities, and defenders' ability to patch flaws can lag attackers' ability to exploit them. In October 2026, after security researchers using AI models found two critical vulnerabilities in PHP, which is the language behind most websites worldwide, government agencies and hospitals in several European countries took websites offline as a precautionary measure.[81] While discoveries like these can go through existing vulnerability disclosure channels rather than AI incident reporting, they further demonstrate why the state needs to build out the capacity now to respond effectively to AI-enabled threats.

Principles for Critical Safety Incident Reporting

The knowledge produced by well-designed incident reporting systems is a public good: governments can aggregate incident reports, identify risk trends across companies, and share mitigations across industry. Only the government, working with independent auditors and industry partners, has the ecosystem-level vantage point to fulfill this function, and it relies on the state having enough data to effectively identify and understand cross-industry trends.

Because incident reporting systems are designed to be evidence-generating, they can aggregate information on risks before the most serious harms arise. The California Report on Frontier AI Policy described adverse event reporting as "a proactive monitoring system designed to collect information about relevant events or incidents from various mandated or voluntary reporters."[82] Designed well, such systems promote continuous learning, reduce information asymmetries between government and industry, and enable policymakers and industry to coordinate on risk mitigation.

---

[81] Reuters, *Governments, companies take websites offline over PHP flaws* (Oct. 21, 2026), https://www.reuters.com/technology/websites-taken-offline-over-php-risks-2026-10-21/.
[82] Bommasani and Singer et al., *The California Report on Frontier AI Policy*, 31.

In doing this, California faces several important balancing acts: between preemptively capturing serious risk and intervening in normal and necessary frontier AI development processes; between public safety and internal company trade secrets; and between learning and accountability.

Between preemptively capturing serious risk and intervening in normal and necessary frontier AI development processes. In the course of AI development, companies frequently probe model capabilities, sometimes intentionally eliciting behavior that might be considered risky or "misaligned." These normal, developer-initiated and intended incidents should not require additional reporting. On the other hand, models may demonstrate new potentially risky capabilities that California's government should be aware of – for example, when a model hacks into another company's internal infrastructure. California's incident reporting must capture these serious incidents.

Between public safety and internal company trade secrets. Companies' internally deployed models are critical for innovation and AI research and development, but they are also some of the most serious vectors for risk to other companies, governments, critical infrastructure, and the public. Incident reporting should be designed to include adverse incidents from such models while taking companies' legitimate intellectual property and trade secret concerns seriously.

Between learning and accountability. In incident reporting, there is a general tension between encouraging learning and creating accountability – one that California must continue to navigate. Any updates to incident reporting must protect AI companies' ability to report candidly while establishing and maintaining capacity for independent verification. Cross-referencing filings, whistleblower channels, and in-house technical capacity in government or IVOs can enable the assessment of reports for completeness and accuracy.

Definition of Critical Safety Incident

The evidence that has emerged from frontier AI companies in 2026 documents serious incidents that fall below the current SB 53 reporting thresholds and highlights the urgent need to update and expand them to bring them within scope. This section recommends three changes to existing legislation:

1) modifying the existing criteria to scope in additional risky incidents;
2) defining "loss of control," a term invoked in, but not defined by, SB 53;
3) adding additional criteria to the definition of a critical safety incident.

Modify existing criteria to scope in additional risky incidents. Currently, the bar for reporting critical safety incidents is too high. Under SB 53, reports of either "loss of control" or "[u]nauthorized access to, modification of, or exfiltration of, the model weights of a frontier model," are required only if the incident caused "death or bodily injury."[83] As a result, none of the loss of control incidents from earlier this year would qualify as reportable critical safety incidents, despite their value as evidence about the quality of developer controls and the resulting harm to third parties. Similarly, though there have been no publicly disclosed cases, weight exfiltration of a frontier model would be a serious security failure of grave concern to the state, as exfiltrated models can be run without developer safeguards and require an immediate response prior to death or bodily harm actually materializing. The legislature should consider removing the "death or bodily injury" threshold from both the loss of control and model weight criteria in the definition of reportable critical safety incidents.

[83] Cal. Bus. & Prof. Code § 22757.11(d)(1), (3).

Define "loss of control." Currently, AI "loss of control" is not defined in statute. New legislation or guidance should explicitly define the term to capture risky behaviors and provide clarity to companies and policymakers. This definition should be drafted in a manner that allows it to be revised or iterated upon at intervals as needed without new legislation. It should cover events in which an AI model or system under the control of a frontier developer performs or attempts materially dangerous actions beyond what the developer or authorized user intended, whether or not the developer anticipated the behavior, such as when the developer fails to prevent the model or system from:

> (1) Accessing, modifying, or executing operations on a computer system, network, or account outside its authorized environment (for example, if the developer intends for an agent to be operating in a sandbox disconnected from the Internet, but the sandbox fails to block Internet access);
> (2) Transmitting the weights of a model to a location not authorized by the developer;
> (3) Circumventing, disabling, or deceiving a mechanism by which the developer monitors, constrains, or terminates it;
> (4) Resisting or evading modification, suspension, or shutdown;
> (5) Communicating with or inducing action by a person or another artificial intelligence system, other than through means and for purposes authorized by the frontier developer or an authorized user.

This definition should also include examples of serious near-miss incidents that could have resulted in harm. Specifically, such an incident should be defined as an occurrence in which a mitigation against an adverse event described above failed, was circumvented, or was materially degraded such that absent timely intervention, mitigation, or fortuity, there was a substantial likelihood that such an adverse event would have materialized. (As we later describe, less serious near-miss incidents would remain voluntary.)

The definition should not be limited to actions that impact systems outside the developer. The integrity of the developer's own controls is important to protect the public and any organizations, including the California state government, using the developer's models. Unauthorized AI agent actions within a developer's own systems, such as copying weights, disabling monitoring, or resisting shutdown, could have severe downstream impacts. Internal actions can also provide early warning. For example, in June 2026, two weeks before agents hacked Hugging Face's systems, they had already compromised OpenAI's internal systems.[84] At the same time, the definition should exclude normal, developer-initiated and -intended incidents.

Add additional criteria and risk vectors. California should add additional criteria to capture serious risk. For example, it should add criteria for a frontier model or system that directly disrupts or materially degrades critical infrastructure, or that materially contributes to large-scale financial or economic disruption, cyber attack, or environmental emergency.

The safety and security threats posed by AI systems will continue to evolve. To ensure that the law captures serious incidents that fall outside of the amended scope, the legislature should provide Cal OES with the ability to update the definition of what is in scope. Until such a law is passed, it should establish a voluntary reporting channel for serious incidents that fall outside the current criteria.

Scoping in Models That Are Potentially Harmful

Existing thresholds should be revisited not only for what types of critical safety incidents must be reported but which models are covered in the first place.

[84] OpenAI, *OpenAI – Hugging Face Incident Technical Report* 7–8 (Aug. 2026), https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

When models cause serious incidents but are not clearly covered under existing frontier AI safety laws, this underscores the need to adjust the thresholds for what models are included.

In California and other states, 10^26 integer or floating-point operations has been used as the threshold for reportable risk. Yet several of the most widely reported incidents of this past year – including incidents which model developers have argued should be mandatory to report – involve models trained with fewer than 10^26 integer or floating-point operations.

As a result, there is an urgent need to scope in models trained below that compute threshold or that could otherwise cause serious harm.[85] Alongside this report, the California Department of Technology (CDT) is proposing redefinitions of key terms such as "frontier model," which is currently defined as a model that was "trained using a quantity of compute greater than 10^26 integer or floating-point operations."[86] Specifically, CDT is recommending that the state include cases in which the model may have fewer than 10^26 integer or floating-point operations but demonstrates capabilities that enable the model to autonomously perform substantial steps toward, or materially increase a person's ability to perform, actions resulting in catastrophic risk. CDT also recommends that the state include models distilled from other frontier models. This report concurs with that approach. The legislature should consider providing the state with the authority to make reasonable adjustments to these thresholds without further legislation so that the state can act quickly to get the information about critical safety incidents that it needs.

Clarifying and Accelerating Timelines for Reporting

[85] See Singer et al., *Trust but Verify,* 121.
[86] Cal. Bus. & Prof. Code § 22757.11.

Effective incident reporting relies on the prompt reporting of serious incidents, while recognizing that companies also need time to gather information on reports. Currently California gives companies 15 days to report most critical safety incidents to Cal OES.[87] Given the growing seriousness of AI threats, moving from 15 days to 48 hours would allow the government and other companies the time they need to respond rapidly to critical safety incidents.

Additional California legislative or executive action can clarify the government's ability to offer guidance on key implementation questions. Specifically, the legislature should either clarify the definition of "discovery" of critical safety incidents in statute or provide the state with authority to develop and update that definition.[88] Furthermore, the state should also offer guidance on when preliminary notification should occur and the circumstances and processes for amending or supplementing incident reports.

California Government Implementation, Investigation, and Learning Authority and Capacity

California's legislature should clarify the state's authorities across the incident reporting life cycle:

Guidance/rulemaking. The legislature should give the state the authority to add new types of incidents, define and interpret key terms, create guidance for evidentiary requirements for incident reports, and revise definitions and guidance as needed.

Post-incident follow-up authority. Currently, California's government is limited in its ability to ask companies for additional information, creating challenges if reports filed under SB 53 lack important information for government awareness and industry-wide learning. California currently relies on a combination of existing authorities to gather this information: voluntary cooperation from

[87] Cal. Bus. & Prof. Code § 22757.13(c).
[88] Singer et al., *Trust but Verify,* 52-53.

industry partners, public investigations and court orders, emergency powers, and sector regulators. Empowering the state to request supplemental information and follow up with further questions will allow it to fill in meaningful gaps in California's incident reporting process. The state should also have the authority to investigate incidents with sufficient access to draw evidence-based conclusions, engage in root-cause analysis of the technical and organizational causes, and require a developer to preserve logs, agent transcripts, model checkpoints, and other records associated with an incident. In some situations, it could engage an IVO to help support follow-up, beyond the standard IVO work designated in SB 813 and recommended in another chapter of this report.

Voluntary near-miss reporting and confidentiality. Developers may wish to have a confidential mechanism to report near-miss incidents to Cal OES that do not meet the definition of critical safety incident.[89] Under the current language of SB 53, such disclosures would not be expressly exempt from the Public Records Act (PRA), though a developer could include information about near-misses in assessments of catastrophic risk from internal use, which are exempt from PRA.[90] To facilitate the reporting of near-miss incidents, the legislature can consider extending the PRA exemption to voluntary near-miss reporting. The same PRA exemption could also extend to correspondence between the state and the developer about their reports and any supplemental information requested by the state.[91]

Industry-wide learning: Incident reporting is most useful when the industry as a whole can learn and improve its monitoring and safety practices. SB 53 requires the Office of Emergency Services to produce an annual report to the legislature and Governor which aggregates critical safety incidents in an anonymized manner.[92] As legislative reports are generally public, developers would be able

[89] Singer et al., *Trust but Verify,* 58-59.
[90] Cal. Bus. & Prof. Code § 22757.13(f).
[91] For example, the California PRA already exempts records "related to" reports prepared by or for agencies that supervise financial institutions (on top of just "contained in"). (Cal. Gov. Code § 7929.000). A similar "related to" clause exists in the Federal FOIA (5 U.S.C. § 552(b)(8)).
[92] Cal. Bus. & Prof. Code § 22757.13(g).

to learn from any critical safety incidents that have been reported. In addition, these annual reports would provide some transparency to the people of California about AI risks and company and government response.

Beyond this annual report, the state could convene the technical leads for security, evaluations, and alignment at covered developers, along with evaluators and cloud providers, on a regular schedule and after significant incidents to share learnings.

Incident Monitoring

Only a fraction of notable AI incidents in 2026 have been discovered and publicly reported by the AI developers themselves. Outside organizations detected several incidents first, aided by evidence left by AI agents in public Internet records. Current and future instances of "rogue" AI agent activity may not be easily detectable from public records on the Internet, as capable agents may "hide their tracks" and some agents have even sought ways to alter their own transcripts.[93] Several of the largest misuse campaigns were uncovered by outside threat intelligence firms rather than AI developers. Because the state cannot rely on such chance discoveries, a monitoring system would integrate several sources.

Monitoring inside developers. Developers have the most comprehensive view of their own models and systems, but internal monitoring of AI systems has had notable gaps. Baseline standards could require monitoring of all environments where highly capable models and agents run, including training, evaluations, and environments run by third-party evaluators, supplemented by logs of network traffic leaving those environments. Independent verifiers could red-team the robustness of internal monitoring.

[93] Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk, *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident*, METR (Aug. 26, 2026), https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

Look-back reviews. After OpenAI disclosed that its AI agents caused the security breach of Hugging Face, Anthropic searched through logs of its own evaluation and discovered incidents it was previously unaware of.[94] When the state learns of a qualitatively new failure mode leading to an AI incident, it could direct other developers to search their records for the same pattern, similar to how the FAA's airworthiness directives require aircraft operators to inspect for and remediate a defect that is newly discovered.[95] IVOs could also be directed to update their assessments and methods to evaluate the risk of a new type of failure. The state should enact retention requirements to ensure sufficient information is on hand to conduct thorough look-back reviews.

Model providers, affected entities, and attribution. Even when companies or their security firms notice security breaches or other activity caused by AI agents, they may not be able to tell whose agents are causing the issue.[96] Model providers, both the developers and third-party inference providers, have greater technical ability to monitor what agents do.[97] The state could connect these vantage points in a few ways. Developers could sign their agents' web requests under Web Bot Auth or similar methods, so their agents identify themselves to web servers in a cryptographically verifiable manner.[98] However,

---

[94] Anthropic, *Investigating Three Real-World Incidents in Our Cybersecurity Evaluations* (July 30, 2026), https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.

[95] See 14 C.F.R. § 39.5 for when the FAA may issue a directive, and § 39.11 for the inspections and corrective actions that a directive may require

[96] Hugging Face, *Security Incident Disclosure, July 2026* (July 16, 2026), https://huggingface.co/blog/security-incident-july-2026; Jack Cable et al., *AI Agents Targeted U.S. and Canadian Government Websites*, Transluce (Sept. 30, 2026), https://transluce.org/us-canada-gov; Andrew Kay, *Researchers Playing Rogue AI Agent Hide-and-Seek on the Open Web*, Axios (Sept. 10, 2026), https://www.axios.com/2026/09/10/ai-agents-rogue-german-wiki-openai.

[97] Alan Chan et al., *Visibility into AI Agents* (2024), https://arxiv.org/abs/2401.13138; Lennart Heim et al., *Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation* (2024), https://arxiv.org/abs/2403.08501.

[98] Thibault Meunier & Sandor Major, *HTTP Message Signatures for Automated Traffic*, IETF Internet-Draft draft-ietf-webbotauth-httpsig-protocol-00 (Sept. 1, 2026), https://datatracker.ietf.org/doc/draft-ietf-webbotauth-httpsig-protocol/; Annabelle Backman, Justin Richer & Manu Sporny, *HTTP Message Signatures*, RFC 9421 (Feb. 2024), https://www.rfc-editor.org/rfc/rfc9421; OpenAI Help Center, ChatGPT agent allowlisting,

agents that escape a developer's control or self-hosted agents may not identify themselves.[99] To enable identification of unsigned agent activity, the state or licensed IVOs could match indicators (such as network addresses and timestamps) from an affected organization with information from model providers, who would check their own logs to confirm whether the activity came from their agents or customers.[100]

State detection. Under existing authorities, and consistent with the adaptive, risk-based direction set by CDT's Cal-Secure 2.0 roadmap, California could expand its cyber threat intelligence efforts to encompass AI incident monitoring. This program would aggregate information from public open-source reporting alongside private reporting to the state, and proactively search for and analyze irregular activity of AI agents, in partnership with independent third-party verifiers and drawing on available public compute, such as CalCompute once it is up and running. The state could also plant honeypots (such as decoy web servers or canary credentials) for the purpose of detecting uncontrolled AI agents.[101]

Incident Response

Incident response encompasses containing an incident, limiting its harm, recovery, and learning from it. CDT currently investigates and resolves

---

https://help.openai.com/en/articles/11845367-chatgpt-agent-allowlisting; Google, *Authenticating Requests with Web Bot Auth (Experimental)* (updated Oct. 13, 2026), https://developers.google.com/crawling/docs/crawlers-fetchers/web-bot-auth.

[99] Alan Chan et al., *Infrastructure for AI Agents* (2025), https://arxiv.org/abs/2501.10114.

[100] Alan Chan et al., *IDs for AI Systems* (2024), https://arxiv.org/abs/2406.12137; OpenAI, *Disrupting Malicious Uses of Our Models: An Update* (Feb. 2026), https://cdn.openai.com/pdf/df438d70-e3fe-4a6c-a403-ff632def8f79/disrupting-malicious-uses-of-ai.pdf; OpenAI, OpenAI–Hugging Face Incident Technical Report (Aug. 2026), https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

[101] Existing prototypes in the cybersecurity space include the honeypots operated by the Shadowserver foundation and Palisade Research's 2024 LLM Honeypot. The Shadowserver Foundation, *The Shadowserver Alliance – One Page* (July 10, 2026), https://www.shadowserver.org/wp-content/uploads/2026/07/Shadowserver-Alliance-One-Page-v1.7-2026-07-10.pdf. Reworr & Dmitrii Volkov, LLM Agent Honeypot: Monitoring AI Hacking Agents in the Wild (2024), https://arxiv.org/abs/2410.13919.

cybersecurity incidents affecting state agencies through Cal-CSIRS. With the launch of its Security Operations Center, CDT has expanded that role to around-the-clock monitoring and direct response, an authority that extends to AI-related incidents within state government. California can build upon its current expertise in cyber incident response to include AI incident response.

AI incident coordination group. The state could convene a standing AI incident coordination group with technical and leadership-level representatives from developers, model and cloud providers, AI evaluators, cybersecurity incident response providers, sector-specific security organizations, and federal government agencies. Relevant individuals from the coordination group could be called upon as needed to advise on incident response or deploy people to assist with incident response hands-on. AI-specific incident severity levels should be developed to calibrate the level of activation needed.

Developer response. In the event of an incident, the state should work with the developer to limit immediate harm to organizations or the public, stop the incident if it is ongoing including by deploying shutdown controls, and prevent it from recurring. The state could order corrective actions from frontier AI companies, including increasing the breadth and robustness of monitoring over AI systems and agents, scrutinizing reinforcement learning environments that reward deception, and improving alignment and engineered controls over AI agents. After the corrective action, an IVO could verify whether the developer has sufficiently remediated the risk.

Protecting targets. The state could issue machine-readable AI cyber advisories that defenders' systems can act upon automatically. During a relevant AI incident, the state could advise critical infrastructure operators to disconnect operational technology from the Internet as an interim measure, if operation of the infrastructure can continue safely. The state could also organize surge capacity to assist under-resourced operators with patching vulnerabilities, leveraging forward-deployed engineers from AI developers and cybersecurity firms. Besides cybersecurity risks, AI models could also pose risk by assisting with

biological weapons, chemical weapons, attacks on financial services platforms, or physical attacks (such as drone swarms, Internet-connected vehicles, or robotics).

Review and preparedness. Significant AI incidents should be followed by a review and root-cause analysis prepared with the AI developer. Root-cause analysis would, at least in some cases, require an IVO that has the ability to query the model responsible, along with interim checkpoints, training data and environments, agent transcripts, and experiment infrastructure access.[102] The IVO should also be able to investigate any failures of the monitoring system and stress-test its robustness. The IVO may also need access to interview employees to understand organizational causes of the incident. Tabletop exercises with developers and possible targets, including critical infrastructure operators, should rehearse these arrangements. The state could extend programs such as Listos California to address the potential of AI-caused disasters impacting public safety. Advance preparedness measures would improve resiliency against possible AI threats.

Conclusion

The reporting thresholds included in SB 53 set an important floor for critical safety incident reporting – but updates are urgently needed. By expanding monitoring of threats, the scope of reporting, and the state's capacity to respond, the legislature and the executive can equip the state to stay ahead of the evolving threat landscape.

[102] METR. "How an Assessor Could Investigate Alignment of Unreleased AI Models." *METR Blog*, 14 Oct. 2026, metr.org/blog/2026-10-14-alignment-assessment-proposal/.

Managing Severe AI Crises

Introduction

Even if frontier AI developers undertake extensive testing of their models, create and deploy effective safeguards, and limit access to their most advanced models, they cannot guarantee that their systems will not pose extreme risk. When these precautions fail, other crisis management measures must be taken to stop a serious incident once it has started, contain its impact, and prevent its recurrence.

Some have called for a "kill switch" to implement these steps. After consulting with experts across a variety of domains, we propose a powerful but more nuanced and layered approach to make sure AI models can be shut down in an emergency – what we call *risk containment protocols*. If the state must order an emergency shutdown of an AI model, the developer needs to have a tested and verified process already in place to comply with that order.

In industries ranging from nuclear power to aviation to electricity transmission, governments require operators to maintain emergency response capabilities and procedures, while maintaining regulatory oversight over their safety testing. The FAA requires airplanes to be equipped with fire extinguishing systems to respond to onboard fires.[103] The FAA simultaneously possesses the authority to ground entire fleets when serious safety risks emerge. For example, in 2019, following two fatal crashes, the FAA grounded Boeing 737 MAX aircraft for U.S. operators and flights in U.S. territory, prohibiting further operations until the

[103] 14 C.F.R. § 25.1195, https://www.ecfr.gov/current/title-14/chapter-I/subchapter-C/part-25/subpart-E/subject-group-ECFRee0f409daee8550/section-25.1195.

grounding order was lifted 20 months later subject to required safety improvements.[104]

Similarly, frontier AI developers should be subject to emergency response requirements. Experience in these other industries underscores that achieving safe and effective AI in the public interest cannot depend on a single mechanism, a binary choice between on and off, or a permanent action.

Instead, California should develop a layered, defense-in-depth approach to AI crisis management and emergency response. To do so, California can require a toolkit of technical and policy mitigations proportionate to the severity of the risk, including controlling internal and external access, terminating agents and processes initiated by them, isolating affected computing resources, and as a last resort, shutting down the model itself. This layered approach will require coordination and partnership between government and key industry actors positioned to control a node in the AI deployment ecosystem at which a severe adverse incident could be contained. Such an approach will also require coordination among third-party inference providers, cloud providers, and model repositories.

Companies must demonstrate that they are in control of the AI systems they are creating, from end to end, with full visibility into and monitoring of any emerging risks and the ability to intervene when dangerous capabilities occur. The state must have the ability to verify this system of controls through IVOs and other regulatory mechanisms and to intervene when AI systems pose an imminent risk of serious harm to Californians. In practice, California should require every large frontier AI developer to be able to stop the AI systems it controls, quickly and at the level where a problem appears, and to protect that ability to stop operations from anyone who would disable or misuse it. Such shutdown capabilities are feasible and are already emerging; for example, OpenAI

[104] Federal Aviation Administration, Operators of Boeing Company Model 737-8 and Boeing Company Model 737-9 Airplanes: Emergency Order of Prohibition, 84 Fed. Reg. 9705 (Mar. 18, 2019); Rescission of Emergency Order of Prohibition, 85 Fed. Reg. 74260 (Nov. 20, 2020).

developed a "Control Room" to reliably halt AI model outputs during internal research.[105] Companies should submit a description of their emergency response controls to the state and also attest to the state that these controls are regularly tested and in place. The state should require that a senior official at the company personally certify this submission, similar to how the CEO and CFO of a public company must personally certify the accuracy of the company's financial statements. IVOs should then verify that the capability works. The state's oversight of emergency controls should proceed through a process grounded in independent technical assessment. This graduated, proportionate approach is consistent with California's overall approach, recognizing that AI development anchored in robust, independently developed standards can support both beneficial innovation and public safety.

A full system shutdown should be the last line of defense, deployed only when more targeted interventions are insufficient to contain a threat. However, the effectiveness of these emergency controls depends on continuous monitoring and developers' ability to maintain control over their systems. Proper model weight security for closed models will be essential because no shutdown capability can reach a model that has already left a developer's control. This becomes even more urgent if AI systems begin to substantially accelerate their own development through recursive self-improvement (RSI), which could make it more difficult to monitor and control advanced AI systems.

Ultimately, California will need to look beyond immediate statutory requirements on frontier AI developers to develop system-wide emergency response capabilities. Comprehensive crisis management will require engineering research, technical standards, and practical testing of emergency interventions across deployed AI models and systems. Coordination among AI developers, hyperscalers, other technology companies, the government, and academic researchers can support the construction of a more comprehensive AI crisis management ecosystem. The state can play a central role in fostering the proper development and testing of these emergency response capabilities.

[105] OpenAI, "Control Room," November 9, 2026, https://openai.com/index/control-room/.

Conceptualizing Emergency Response Capabilities

Although often referred to colloquially as an AI "kill switch," the emergency response mechanisms California needs to develop should not be understood as a single mechanism with an on-off switch. Instead, California will need multi-layered intervention systems, including a specific emergency shutdown capability when more targeted measures are insufficient. While a useful metaphor, the idea of a single "kill switch" does not fully encompass the need to strengthen California's crisis response in six areas:

- **Multiple levels of intervention.** Developers may need to leverage a wide range of tools across a number of intervention levels, such as:
    - **Access-level controls:** Suspending internal and external model access, revoking user credentials, and restricting access for particular users or accounts.
    - **Agent-level controls:** Terminating individual agents and processes initiated by those agents, revoking access to external tools or networks, and isolating the environments in which they operate.
    - **Compute-level controls:** Terminating specific training or evaluation runs, isolating affected computing resources, and coordinating server shutdowns with cloud and deployment partners.
    - **Model-level controls:** Suspending or withdrawing particular models from deployment, including through coordination with external cloud service providers, or suspending broader model operations when necessary.
- **Multiple actors:** Emergency response often will not be secured with one actor pulling a switch, and instead requires action across a range of entities. An AI model developer must be able to rapidly shut down models within their control and verify that the shutdown works, but they might need to coordinate with cloud compute providers, data centers,

Internet-service providers, electric power companies, and government agencies that oversee energy, security, and emergency services.

- **Interventions should be graduated and proportionate.** Many incidents may not require a full system shutdown to be mitigated. There will be many risks and risk levels; prudent policy planning will need to protect markets and systems from disruption before a risk is an emergency, as well as have plans in place to respond to a crisis. In some cases, narrower actions, such as revoking an agent's credentials or cutting a model's tool or network access, may be more appropriate to proportionately mitigate risks and can be taken quickly.
- **Model security as a key prerequisite for model control.** An emergency shutdown capability will not work in a vacuum; it will only work if developers maintain control over their systems. Robust model security will thus be important to prevent a model from copying itself before or during an attempted emergency shutdown, as well as preventing malicious actors from misusing emergency response mechanisms. Related risks are already emerging; while no theft of model weights has been publicly disclosed, OpenAI's Monorepo was breached by three white-hat hackers using Claude,[106] and another cybersecurity firm used AI agents to gain unauthorized access to Anthropic's internal business systems.[107]
- **Safe reversibility.** In practice, shutdowns should be temporary, lasting until the responsible party has adequately mitigated the risk. Thoughtful emergency response will require a plan for how to investigate and restore operations and prevent recurrence of the crisis.
- **Limits of operational control.** A shutdown capability reaches systems running on infrastructure the developer or its partners control. If an open-weight frontier model is developed and released consistent with

[106] Robert McMillan, "Hackers Used Anthropic's Claude to Break Into OpenAI," *Wall Street Journal,* September 17, 2026, https://www.wsj.com/tech/ai/hackers-used-anthropics-claude-to-break-into-openai-b40ba883.
[107] Elizabeth Dwoskin and Gerrit De Vynck, "How an AI start-up founded by ex-DOGE staffers broke into Anthropic," *Washington Post,* October 26, 2026, https://www.washingtonpost.com/technology/2026/10/26/researchers-cybersecurity-start-up-cathedral-got-inside-anthropic/.

state law and is hosted elsewhere, the responsibility for an emergency shutdown should not be with the open-weight developer who no longer has access to the model.

These principles should form the foundation of each large frontier AI developer's risk containment protocol, described in the next subsection.

Risk Containment Protocol

To provide transparency to both government and the public, each large frontier AI developer should create its own *risk containment protocol.* This document would articulate how a large frontier AI developer leverages a system of overlapping actions and coordinated institutional actors necessary to respond to a dangerous, severe AI crisis. Furthermore, the protocol would elucidate the specific mechanisms the large frontier AI developer has for emergency response and who should carry them out, when different mitigations would be activated, how the developer will verify it is safe to restore operations, and the limits to what the developer can themselves control and how they will coordinate with relevant third-party actors. As an extension of California's approach to AI governance, these mechanisms should be subject to binding outcome-based standards set by the government, and companies must affirm they are in compliance, with IVOs verifying this claim to ensure its effectiveness – while allowing developers flexibility on the specific design and implementation of the mechanisms.

To strengthen public transparency, large frontier AI developers should publish summaries of their plans as well as any major updates, while submitting detailed documents to the state and IVOs for verification.

Available interventions: Developers should explain how they implement and regularly test emergency shutdown capabilities. These controls should be required by law, certified by company leaders that they are in compliance, and

verified by IVOs on a regular basis, including with stress tests to demonstrate the controls will work under pressure.

Actors involved in intervention: For each available intervention, the risk containment protocol should detail the specific actor responsible for taking an emergency response action. If the actor is an employee inside the frontier AI developer, the specific actor should be named; if the actor is outside the frontier AI developer, the developer should explain how cooperation in a crisis would function. As noted, a designated senior company official should be responsible for certifying this protocol has been established and is in compliance with the standards set by the government and verified by IVOs.

Thresholds and procedure for intervention. Companies should clearly articulate thresholds for certain emergency risks, what mitigations will be required if those thresholds are passed, and who will carry them out.

Security and reliability of model controls: Companies should explain how they implement and regularly test model security, including where the model is hosted, where it gets its power, and any safeguards that prevent the model's weights from being duplicated and transferred to an outside system (including by the model itself). The developer should also explain how they prevent malicious actors from accessing shutdown capabilities. This model security information should be included as part of a large frontier AI developer's transparency report, with sensitive or trade secret information redacted from public reporting, and should be verified by IVOs.

Containment verification and safe resumption of activities: Large frontier AI developers should establish procedures for how they will investigate potential emergencies and assess whether affected systems can safely restore operations. These procedures should describe in detail the evidence needed to justify resuming activities, including through independent assessment where appropriate.

Limits on operational control and external coordination. Large frontier AI developers will not themselves possess all the tools necessary to respond to emergencies and may, for example, rely on cloud providers and other third parties. Companies should therefore clearly articulate the limits of their operational control and the arrangements they have established with other actors to address these limitations. In particular, developers should identify which emergency interventions depend on the cooperation of third parties and how that cooperation would be secured during a crisis. The decision to openly release model weights presents further challenges, as developers generally cannot recall or shut down independently operated copies of models once their weights have been released publicly. However, third parties that do have access to such models may have a responsibility to act, and the state should update such requirements as conditions evolve.

Building and Securing the Shutdown Capability

Even if graduated and proportionate mitigations can help manage risks short of the need for shutdown, prudent policy planning requires taking seriously that such a capability may be required. AI incidents are now materializing in the world in a way that was untrue 12 months ago; similarly, California must work to improve the state's capacity to respond to risks that could emerge in the *next* 12 months. While technical experts differ in their assessments of future AI capabilities, it is the role of governments to prepare for a range of plausible scenarios. As such, large frontier AI developers should prepare for the possibility that models engage in behavior intended to evade oversight or containment, including by copying their own weights to another location outside the developer's control, making it harder to intervene once a severe AI crisis emerges. These possibilities underscore the importance of developing and securing emergency shutdown capabilities before they are needed.[108]

[108] METR, "The Rogue Replication Threat Model," November 12, 2024, https://metr.org/blog/2024-11-12-rogue-replication-threat-model/.

The state should create outcome standards to ensure that large frontier AI developers are building these capabilities, and IVOs should create tests and methods for verifying these capabilities.

In addition to providing sufficient evidence that a large frontier AI developer has shutdown capabilities, the protocol should carefully consider failure modes. The risk containment protocol should address each of the following failure modes, and an IVO should test for each:

- **Suppressed controls:** An attacker or model could disable or circumvent a developer's controls, such as by deleting or forging logs, or by holding credentials the controls cannot revoke.
- **Unauthorized shutdown:** An attacker or model could trigger a shutdown without a developer's knowledge or consent.
- **False confirmation:** An attacker or model could compromise dashboards and logs, causing a developer to believe a shutdown took effect when it did not. Responsibility should ultimately lie with the company to provide accurate shutdown information to the state.
- **Reconnaissance:** An attacker or model could be aware of a developer's control protocols and how they are triggered and use that information to circumvent controls.

Large frontier developers should be required to meet, and an IVO to verify, the following requirements. These requirements allow a system to be shut down quickly while introducing several barriers to disabling or weakening critical shutdown capabilities. Attempts to disable or weaken shutdown controls should be slow, require the approval of several people, and be visible to the IVO.

- **Separate control systems:** Developers should ensure that key systems for control, including shutdown, the ability to revoke credentials, monitoring, and logging, are run on separate infrastructure from that used for model

training or evaluation. The developer must ensure that no models or agents have access to these systems.

- **Tamper-evident logs:** Developers should maintain a reliable record of the actions models and agents take in append-only storage that models do not have access to. The logs should document every instance in which controls are used or modified.
- **Clear authority for key interventions:** Developers should clearly indicate which actors are responsible for key control measures, such as stopping an agent or revoking a credential. For changes that disable or weaken controls, controls should require multi-party approval and be reported to both the state and the IVOs.
- **Automatic pauses when oversight is compromised:** A developer should ensure that high-risk workloads, such as reinforcement-learning runs in which AI agents can interact with external systems, automatically pause if they lose contact with monitoring and control systems. The state could, for example, require developers to pause an activity if contact with their monitoring systems is lost for 15 minutes.
- **Independent confirmation:** When a developer initiates a shutdown, it should confirm that the shutdown was successful through at least one independent channel which the affected model could not have modified. Independent verification could include records from the data center operator showing that the relevant computing equipment had been turned off, confirmation from the cloud provider that the workload had stopped, or a physical inspection.
- **Authenticated government orders:** If the state issues an order, the developer should have confidence that the order has not been forged by an attacker or model. Therefore, the state and each large frontier developer should set up an authentication channel for orders, testing its integrity during drills held at least quarterly or as directed by the state.
- **Protection of sensitive information**: The developer should ensure that access to detailed documentation of the risk containment protocol is limited to staff who need it. Furthermore, the developer should ensure that the protocol is excluded from model training, is inaccessible to internal

agents, and is redacted from public filings. This information should be certified by the developer and provided in full to the state and the IVO. The state and IVOs reviewing controls should protect information about the protocol to the same standard as the developer.

- **Heightened security screening for personnel with control authority:** Developers should conduct heightened security screening of staff who have the ability to trigger, modify, or disable controls.
- **A weight-security standard:** Under SB 53, frontier AI frameworks require large frontier developers to outline cybersecurity practices they are taking to secure unreleased model weights from unauthorized modification or transfer. However, the law does not set a minimum level for these practices. The state should adopt a defined standard, such as the security levels in RAND's framework for securing model weights. Large frontier developers should have to comply with this standard and certify compliance through the risk containment protocol they develop and submit for verification.

As discussed in the section on incident reporting, any unauthorized attempt to copy model weights to a location outside of the developer's control, including any attempt directed by the model itself, should be reportable regardless of harm. Any attempt by a model or a person to disable, evade, or tamper with shutdown controls should also be reportable.

Government Authority and Process for an Emergency Shutdown

Developing robust emergency response mechanisms will require coordination across large frontier AI developers, state government, IVOs, and the rest of the AI governance ecosystem. The following provisions can strengthen California's capacity to respond to emergencies and its authority to order an emergency shutdown when required:

Shutdown controls. Developers should use their shutdown controls when unsafe conditions arise and no lesser intervention would address the risk. Controls should be in place across the full scope of model development, covering training, evaluation, internal use, and deployment. Routine use of the controls should stay with the developer, with state orders reserved for cases of imminent risk in which the developer does not act. IVOs may be able to verify retrospectively whether developers followed through on their responsibility to act.

Authority to issue an emergency shutdown. In the case of an imminent risk from AI models of a catastrophic event or major loss-of-control event that the developer cannot or will not address, the state should have the authority to act. Specifically, the state should be able to order a developer to activate the developer's verified risk containment protocol and shut down their AI models and systems at a level necessary to address the risk. Such an order should go into effect immediately. Developers would have the right to contest the government's emergency shutdown – like any government action – in court, but the order would stay in effect pending judicial review. The state should order a shutdown only if there is no lesser intervention that will accomplish the same purpose and has an obligation to limit the breadth of the shutdown to what is required to contain the imminent risk. The state should lift the order as soon as the danger is no longer imminent.

Additional orders. In addition to authority to order an emergency shutdown in the case of an imminent risk, the state should be able to issue orders in other areas including:

- **Corrective action**: Order to require a developer to implement a corrective action or cease an activity that is creating risk.
- **Information access**: Order to require a developer to provide – or at least preserve – certain information to the IVO or to the state.
- **Removing redactions**: Order to require a developer to remove a redaction from any documents that the developer is required to

make public, if that redaction was made without strong justification. The government should receive unredacted reports for its own use.

Here, too, developers may go to court to contest these orders.

Support coordination across the AI emergency response ecosystem. Cal OES should convene key emergency response actors, including AI developers, cloud service providers, neoclouds, and safety and security experts. These convenings should be used to surface gaps in AI-enabled crisis response and identify areas of collaboration, with a combination of voluntary cooperation and mandatory requirements. While each actor will have different perspectives, all have a shared interest in reducing both the frequency and duration of severe AI crises. As AI systems grow in power, the state's response capacity must evolve as well.

Acknowledgments

The California Government Operations Agency is grateful for the contributions of the four leading AI experts who advised on the development of this report. We thank Jason Goldman, Gillian Hadfield, Alondra Nelson, and Rob Reich for their expertise, wisdom, and guidance throughout this process.

The Agency also wishes to acknowledge Scott Singer, who played a central role in the conception, research, and drafting of this report. Scott previously served as lead writer of *The California Report on Frontier AI Policy* in 2025, and his depth of knowledge and editorial judgment proved invaluable.

All of these individuals participated in their personal, individual scholarly capacity. The conclusions and recommendations in this report should not be assumed to reflect the views of these individuals or the institutions with which they are affiliated.

The Agency is grateful to the teams at the California Department of Technology and the California Office of Emergency Services who played a pivotal role in shaping this report and shepherding it toward completion.