INSIGHTS

Artificial Intelligence

California Enacts Broad AI Safety Measure Mandating Standardized Disclosure and Transparency Practices for Developers of Large AI Models

Supporters see needed safeguards for frontier AI models while critics warn of significant compliance burdens

By K.C. Halm and Alexander Sisto

10.02.25

Key Takeaways

- Latest California AI law mandates heightened transparency, disclosure, and reporting obligations for developers of large AI "frontier" models. Such developers must publicly disclose AI governance practices and guardrails in "AI frameworks," publish "transparency reports" concerning key features of new models, and report certain "critical safety incidents" to state agencies.

- Measure codifies elements of a recent California expert working group report on AI safety which recommends heightened transparency and disclosure obligations to regulate developers of AI models.

- Legislature is concerned that large models have "capabilities that pose catastrophic risks from both malicious uses and malfunctions, including artificial intelligence-enabled hacking, biological attacks, and loss of control."

- Transparency and disclosure duties are buttressed by provisions of the new law that protects whistleblowers who disclose to regulators or employers dangers to public health or safety resulting from a catastrophic risk, or violations of the law regarding large AI model behavior.

- Measure establishes a public cloud compute cluster, "CalCompute" at the University of California, that will provide AI infrastructure for startups and researchers.

On Monday, September 29, California Governor Newsom signed a new AI law in California mandating significant new disclosure, reporting, and transparency obligations for developers of large AI models. The measure,

known as the Transparency in Frontier Artificial Intelligence Act (SB 53), requires certain developers of large AI "frontier" models to: (i) proactively disclose AI governance and risk mitigation practices in an "AI framework" report; (ii) adopt practices to ensure greater transparency in defining and assessing catastrophic risk thresholds arising from potential uses of large AI frontier models; and (iii) report certain AI safety incidents. The bill will become effective **January 1, 2026**.

This measure adopts key elements of a California expert working group report (the California Report on Frontier AI Policy) issued earlier this year on the development of AI safety guardrails, and mirrors a similar measure approved by the New York legislature, which is now pending before New York Governor Hochul.

The author of SB 53, California State Senator Scott Wiener, is the author of a more fulsome AI regulatory proposal offered in a previous legislature that was ultimately vetoed by Governor Newsom following the 2024 legislative session. This new measure reflects Senator Weiner's attempt to step back from direct regulatory mandates regarding operations and instead rely on increased transparency and reporting duties to achieve similar goals.

Disclosure of AI Governance Practices in "AI Framework"

The measure directs "large frontier developers"—entities that (1) develop AI models trained using a computing power greater than 10^26 integer or floating-point operations (definitions to be updated annually), and (2) have gross annual revenues in excess of $500 million—to disclose on their website a "frontier AI framework" that describes the company's AI governance, safety and security practices. Such disclosure must describe, among other things, how the company:

- Incorporates national and international governance standards, and

industry-consensus best practices into the frontier AI framework's governance procedures;

- Defines thresholds to identify and assess whether a frontier AI model is capable of causing "catastrophic risks";
- Uses mitigation strategies to address the potential for catastrophic risks;
- Implements cybersecurity practices to secure unreleased model weights from unauthorized modification or transfer by internal or external parties; and
- Responds to critical safety incidents.

Frontier AI framework reports mandated by this measure must be updated at least once a year, or whenever the frontier model developer undertakes a "material modification" to the framework.

Transparency and Disclosure Obligations

All frontier model developers—i.e., not only those that have over $500 million in annual revenue—are also required to disclose on their website certain information about the frontier model in a so-called "transparency report." Public reports must disclose model release dates, supported languages, modalities of output, intended uses of the model and generally applicable restrictions or conditions on use of the model. Notably, this mandate differs from similar obligations under the EU AI Act, which requires disclosures directly to regulators.

Transparency reports must also include assessments of "catastrophic risks from the frontier model conducted pursuant to the large frontier developer's frontier AI framework." Such reports must disclose the results of the assessments, involvement of third-party evaluators (if any), and any steps taken to fulfill the steps of the company's frontier AI framework.

Frontier AI developers that already disclose this information in system or model cards would be deemed in compliance with these transparency obligations.

SB 53 also requires *large* frontier developers to disclose to a state agency "any assessment of catastrophic risk resulting from internal use of its frontier models" every three months, or other reasonable timeframe. The measure specifically prohibits any materially false or misleading statement about catastrophic risk, or a company implementation of, or compliance with, its frontier AI framework.

Reporting of "Critical Safety Incidents"

A government agency will establish a new reporting mechanism for frontier developers or members of the public to report "critical safety incidents," defined as: (1) the unauthorized access to, modifications 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; or (3) loss of control of a frontier model causing death or bodily injury. Large frontier developers are required to report such incidents within 15 days of discovery, unless the incident is deemed to pose an imminent risk of death or injury, which triggers more immediate reporting obligations to appropriate authorities. AI critical safety incident reports must include the nature of the incident, date of occurrence, and whether the incident was associated with the use of a frontier model.

SB 53 authorizes the state Attorney General, or the agency authorized to implement the reporting process (the Office of Emergency Services), to share any submitted critical safety AI incident report with the Legislature, Governor, the federal government, or "appropriate" state agencies.

These reports are exempt from the California Public Records Act, and state agencies are directed to be mindful of risks of disclosure of trade secrets, public safety, cybersecurity, and other risks arising from the disclosure of

critical safety AI incidents.

The new measure authorizes the Office of Emergency Services to adopt regulations designating one or more federal laws, regulations, or "guidance documents" that are substantially similar to, or more rigorous than, the standards under SB 53 to act as a safe harbor if larger frontier AI model providers can demonstrate compliance with such federal law, rule, or guidance.

Whistleblower Protections

SB 53 also provides certain protections for whistleblowers, including prohibiting frontier developers from making, adopting, enforcing, or entering into a rule, regulation, policy, or contract that prevents a covered employee—defined as an employee responsible for assessing, managing, or addressing risks of critical safety incidents—from disclosing information related to those incidents, or retaliates against such employee for disclosing such information to the Attorney General, a federal authority, a person with authority over the covered employee, or another covered employee who has authority to investigate, discover, or correct the reported issue.

The measure also requires that large frontier developers establish internal processes through which a covered employee may anonymously disclose information to the developer if the covered employee, in good faith, indicates that the large frontier developer's activities present a specific and substantial danger to the public health or safety resulting from a catastrophic risk or that the large frontier developer violated the statute.

Violations of the transparency and disclosure elements of the new law will be subject to civil penalties "in an amount dependent upon the severity of the violation" but not to exceed $1 million per violation. The measure also establishes procedures to enforce protections and authorizes the collection of attorneys' fees for any plaintiff who brings a successful action

for a violation of the measure.

Public Cloud Computing Consortium

This new measure also authorizes the development of a "public cloud computing cluster," which is intended to expand access to computing resources necessary to develop and train large models and foster research and innovation that "benefits the public." This initiative, labeled CalCompute, will be operated by the University of California and is intended to lead to the development of a hosted cloud platform and human expertise to support, train, and facilitate the use of the platform.

Related Insights

07.28.25
INSIGHTS

The AI Mandate: Trump Administration's Executive Orders and AI Action Plan

06.27.25
INSIGHTS

The Generative Slate: Two Courts Find Fair Use in GenAI Training

04.23.25

INSIGHTS

Utah Enacts Multiple Laws Amending and Expanding the State's Regulation of the Deployment and Use of Artificial Intelligence