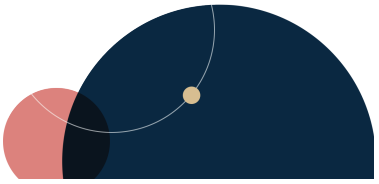
Steptoe

STEPTECHTOE | OCTOBER 8, 2025

California Is Instituting New Compliance Obligations Under the First AI Safety Act To Go Into Effect in the United States

Authors

Christian M. Auty, Tyler Evans, Michel Paradis, Jennifer Quinn-Barabanov, Claire Rajan, Christopher Suarez, Madeline Carr, Peyton Thomas



Overview

On September 29, 2025, California Governor Gavin Newsom signed into law Senate Bill 53, the Transparency in Frontier Artificial Intelligence Act (TFAIA), which is the first state law to directly target large-scale generative AI platforms with safety and transparency requirements. TFAIA emerged after a series of more restrictive AI proposals introduced earlier in the legislative session and reflects the work of the Governor's Frontier AI Working Group. Although groundbreaking, the law ultimately adopts a narrower set of requirements than initial proposals, which would have included additional provisions mandating a "kill switch" for certain AI and third-party audits.

California has already enacted several laws regulating AI, including: AB 1836, Prohibiting Digital Replicas of Deceased Personalities (effective Jan. 1, 2025); SB 942, California AI Transparency Act (effective Jan. 1, 2026); AB 2602, Contracts against public policy: Digital Replicas (effective Jan. 1, 2025); SB 11, Artificial Intelligence Technology (effective Jan. 1, 2026); AB 489, Healthcare Professions: Deceptive Terms or Letters (Artificial Intelligence) (effective Oct. 1, 2025); AB 853, Building on the California AI Transparency Act (effective Jan. 1, 2027). Together, these measures signal California’s continued intent to establish a leading role in AI governance, transparency, and safety. Key compliance obligations and recommendations for companies to prepare are outlined below.

The TFAIA acknowledges the possibility for federal safety and transparency standards and accounts for that possibility with regard to the transparency obligations it sets forth. At the same time, however, the federal government's AI Action Plan has threatened action against states that pass AI requirements that are perceived to be burdensome or restrictive of innovation. Although currently unlikely to pass, members of Congress also continue to introduce legislation that would preempt state regulation of AI.

What the New Law Does

Effective January 1, 2026, TFAIA applies to large frontier developers, which have or are in the process of training a foundation model using computing power greater than 10^26 floating-point operations if the developer—and its affiliates, potentially including major investors—combined had revenues exceeding $500 million in the preceding fiscal year. Foundation models are defined as being trained on broad data sets, designed for generality of output, and adaptable to a wide range of distinctive tasks. The law requires covered entities to clearly and conspicuously publish a "frontier AI framework" on their websites, explaining how they assess and mitigate risks, and how the company incorporates national, international, and industry-consensus standards. The law also establishes various mechanisms for reporting safety issues, particularly regarding what the statute calls "catastrophic risks," with developers being required to report on incidents within 24 hours to 15 days depending on severity, and includes protections for whistleblowers.

This is the first state legislation to impose transparency and reporting requirements on large-scale generative AI platforms, many of which are based in California. Although Colorado's AI Act also requires transparency and mandatory reporting, it will not take effect until June 2026, six months after TFAIA's effective date, and that legislation is subject to possible revision at a later date.

The closest existing legislation to the TFAIA can be found in the EU AI Act's provisions and requirements for General Purpose AI Models (GPAI). Companies thinking about compliance, should therefore consider both TFAIA and the EU AI Act in tandem.

What's In the Law

Broadly, the TFAIA:

- Requires publication of frontier AI frameworks to explain how developers identify and respond to "critical safety incidents,"[1]which must be updated at least once per year.
- Requires adoption of a new reporting mechanism frontier AI companies and the public can use to report critical safety incidents to California’s Office of Emergency Services (COES).
- Requires that transparency reports are made publicly available, at or before deployment of a new or substantially modified model, that summarize risk assessments, results, the role of third-party evaluators, and other compliance measures. Trade secrets and sensitive information may be redacted from these reports if accompanied by a justification, and COES is mandated not to transmit any information to state legislators, governors, or other public bodies that reveal trade secrets or create a cybersecurity or other public safety risks.
- Establishes a new consortium to create "CalCompute" within California's Government Operations Agency to advance the development and deployment of "safe, ethical, equitable, and sustainable" artificial intelligence. The consortium must deliver a framework for CalCompute by January 1, 2027, so that the state has a mechanism to internally test and research AI model development.

- Requires adoption of whistleblower protections for certain covered employees who disclose significant health and safety risks posed by frontier AI models, so long as those employees have "reasonable cause to believe" they have information that bears on specific and substantial dangers to public health or safety. Developers must also establish and implement anonymous, internal reporting processes. Disclosures and responses related to these processes must be shared with directors and officers at least quarterly.
- Establishes civil penalties for noncompliance of up to $1 million per violation of TFAIA's transparency, reporting, or risk requirements, which are enforceable by the California Attorney General.
- Directs the California Department of Technology to annually recommend appropriate updates to TFAIA based on multistakeholder input, technological developments, and international standards.

Implications

The implications for frontier AI model developers are substantial. Covered companies will need to consider developing or strengthening internal safety risk assessment and documentation processes, ensuring that safety and security frameworks are up to date and in compliance with global AI standards and industry best practices. If they do not, they will need to publicly report how they fail to meet listed criteria. In addition, regardless of how they decide to proceed with safety risk assessment and documentation processes, they will need to develop public notices and establish incident-response procedures. Failure to comply with the transparency, reporting, or risk requirements could result in enforcement actions by the California Attorney General, along with potential risks of losses of goodwill amongst the AI-using public.

The California AG may seek civil penalties of up to $1 million per violation under TFAIA. Examples of potential violations include failing to publish or update the required frontier AI framework, failing to make the mandated transparency disclosures, or failing to submit critical safety incident reports. The Act also creates a private right of action for internal whistleblowers subject to retaliation or impermissible reporting restrictions, allowing them to seek injunctive relief and recover attorney’s fees if they prevail and shifts the burden of proof to developers once an initial showing is met. The Act does not include a private right of action for enforcement of its other provisions.

How Frontier AI Developers Should Prepare

To mitigate risk and demonstrate compliance, frontier AI developers subject to TFAIA should move quickly to evaluate their current practices and identify gaps that could expose them to liability or unwanted scrutiny under the new law. Considerations for companies that run large-scale generative AI platforms include:

- **Develop a Frontier AI Framework.** Create and maintain a public-facing framework that explains how the company mitigates, identifies, and manages critical safety risks for all its deployed, public-facing AI models, with clear ties to industry standards and best practices.
- **Develop Procedures to Revise Frontier AI Frameworks and Publish Transparency Reports on an Ongoing Basis.** Review and update frontier AI frameworks at least once per year and publish transparency reports upon release of (or substantial changes to) any frontier model.
- **Assess Whistleblower Procedures.** Ensure internal reporting channels allow for anonymous submissions by employees and that Directors & Officers receive regular updates (at least quarterly) as required under TFAIA.

- **Strengthen Incident Response Protocols.** Develop clear systems to promptly detect, document, and report critical safety incidents to California's Office of Emergency Services.
- **Document the Processes.** Keep detailed records of risk assessments and testing to support transparency required for the deployment of new or substantially modified models.

The TFAIA takes effect on January 1, 2026. Steptoe will continue tracking regulatory developments across federal agencies and legal developments at the state level through Steptoe's AI Legislative Tracker. Please contact members of our interdisciplinary artificial intelligence team spanning our Commercial, Consumer and Government Litigation, Energy, Financial Innovation and Regulation, Intellectual Property, International Trade and Regulatory Compliance, Investigations and White Collar Defense, and Telecom and Technology practices to assess how AI regulations may affect your operations or investments.

---

[1]Critical Safety Incident is defined to include any of the following: "(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."

Practices

AI, Data & Digital

Artificial Intelligence