Davis Polk

California governor signs Transparency in Frontier Artificial Intelligence Act

October 9, 2025 | Client Update | 7-minute read

The Transparency in Frontier Artificial Intelligence Act is the first U.S. law focused on developers of frontier AI models and will go into effect on January 1, 2026.

Background

On September 29, California Governor Gavin Newsom signed into law the Transparency in Frontier Artificial Intelligence Act ("TFAIA"), which will go into effect on January 1, 2026. Among other things, TFAIA requires "frontier developers"—in effect, developers of general-purpose AI models that were trained using high levels of computational power—to publish a framework describing how they incorporate best practices and relevant benchmarks, identify and mitigate risks, respond to critical safety incidents and institute internal governance practices. TFAIA also requires frontier developers to disclose the results of "catastrophic risk" assessments before deploying new or substantially modified frontier models, and to notify California's Office of Emergency Services of any "critical safety incident" within a certain number of days, depending on the severity of the incident. TFAIA also includes robust whistleblower protections.

TFAIA follows Governor Newsom's veto of the Safe and Secure Innovation for Frontier Artificial Intelligence Systems Act, also known as S.B. 1047, in September 2024. The earlier bill would have required AI developers to conduct rigorous pre-deployment safety checks, required AI developers to implement a "kill switch" and mandated annual third-party audits of AI models. Governor Newsom explained his veto by noting that S.B. 1047 failed to consider whether an AI system "is deployed in high-risk environments, involves critical decision-making or the use of sensitive data[,]" and thus "could give the public a false sense of security[.]" Following his veto, Governor Newsom convened the Joint California AI Policy Working Group to draft recommendations on the safe development of frontier AI. TFAIA builds on that panel's recommendations.

In his signing statement, Governor Newsom touted TFAIA's new public safety and transparency requirements, but also noted "the absence of a comprehensive federal AI policy framework and national AI safety standards[,]" and signaled that TFAIA should be revisited in the event such federal legislation is enacted.

Applicability

TFAIA applies to "frontier developers," defined as a person who has trained or initiated the training of a frontier AI model—a general-purpose AI model that was trained using a sufficiently high threshold of computing power. However, certain of TFAIA's requirements apply only to "large" frontier developers—meaning, frontier developers that, together with their affiliates, had annual gross revenues in excess of $500 million.

Notably, TFAIA is the first U.S. legislation to define covered frontier models by specifying a minimum threshold of computing power ("compute") as a proxy for the model's sophistication and therefore its risk. This mirrors the EU AI Act, which also utilized a compute threshold to define general purpose AI systems, though TFAIA's threshold (10^26 integer or floating-point operations) is an order of magnitude greater. Under TFAIA, this includes compute from original training runs, subsequent fine-tuning, reinforcement learning and any other "material modifications" applied to a foundation model. TFAIA provides that, starting in 2027, the California Department of Technology must annually reassess the definitions of "frontier model," "frontier developer" and "large frontier developer," to ensure they remain current.

TFAIA adopts the Organisation for Economic Co-operation and Development's ("OECD") definition of "artificial intelligence model"—an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs.

Disclosure requirements

Frontier AI frameworks

TFAIA requires large frontier AI developers to publish on their websites a "frontier AI framework" that describes how the developer approaches specific aspects of developing frontier models, and prohibits misrepresentations about these practices. These frameworks must be updated annually, and within 30 days of any material modification. However, TFAIA is not just a disclosure-based regime—it requires large frontier developers to implement and comply with any practices described in their published framework, thereby creating an additional source of compliance risk for developers.

Required topics for a frontier AI framework include how the developer:

- Incorporates national and international standards, as well as industry best practices;
- Defines thresholds used to assess whether a frontier model could present a catastrophic risk, and manages these risks, including risks resulting from a frontier model circumventing oversight mechanisms;
- Applies mitigations to address the potential for catastrophic risks, and reviews assessments and mitigations as part of any decision to deploy or use such a model, even internally;
- Uses third parties to assess the potential for catastrophic risks and the adequacy of mitigations;
- Revisits and updates its frontier AI framework, and approaches the circumstances that trigger updates and/or disclosure requirements;
- Deploys cybersecurity practices;
- Identifies and responds to "critical safety incidents"; and
- Designs and implements internal governance practices.

Catastrophic risk disclosure requirements

TFAIA also requires large frontier developers to disclose the results of catastrophic risk assessments before deploying new or substantially modified frontier models, and prohibits materially false or misleading statements about catastrophic risk.

TFAIA's disclosure requirements regarding catastrophic risk are triggered where there is a foreseeable and material risk that a frontier developer's development, storage, use or deployment of an AI model will materially contribute to the death of, or serious injury to, more than 50 people, or more than $1 billion in property damage arising from an AI model doing any of the following:

- Providing expert-level assistance in the creation or release of a chemical, biological, radiological or nuclear weapon;
- Engaging in conduct with no meaningful human oversight, intervention or supervision that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion or theft, including theft by false pretense; or
- Evading the control of its frontier developer or user.

Covered developers are also required to send the California Office of Emergency Services ("OES") summaries of any catastrophic risk assessments every three months, or on some other "reasonable schedule" specified in writing to the OES.

"Critical safety incident" disclosure requirements

Moreover, TFAIA requires covered developers to notify OES of any critical safety incident within 15 days of discovering the incident, or within 24 hours if the incident presents an "imminent" risk of death or serious injury. A "critical safety incident" is defined as:

- Unauthorized access to, modification of or exfiltration of the model weights of a frontier model, and which results in death or bodily injury;
- Harm resulting from the materialization of a catastrophic risk;
- Loss of control of a frontier model causing death or bodily injury; or

— A frontier model that uses deceptive techniques against the frontier developer to subvert the controls or monitoring of its frontier developer in a manner that demonstrates materially increased catastrophic risk.

TFAIA also calls for the creation of a mechanism through which members of the public—and developers, confidentially—may report critical safety incidents to the government. Like the mandatory reports described above, these reports would also be exempt from California's information access laws. However, beginning in 2027, OES will publish anonymized summaries of reported critical safety incidents.

Governance and whistleblower protections

TFAIA also contains whistleblower protections. Among other things, TFAIA requires frontier developers to provide anonymous internal reporting mechanisms to employees, and prohibits retaliation against employees who are responsible for assessing, managing or addressing the risk of critical safety incidents.

In addition, TFAIA prohibits employment contracts that prevent employees from making public safety reports, and requires companies to notify employees of their whistleblower rights. Employees must receive this notice annually, and companies must retain records of employees' acknowledgments.

Enforcement

TFAIA includes two enforcement regimes. Failure to comply with TFAIA's reporting and disclosure requirements, or with published frontier AI frameworks, may result in a financial penalty of up to $1 million per violation. California's Attorney General has exclusive authority to bring actions for such violations.

TFAIA's whistleblower protections, however, may be privately enforced by individual employees through civil lawsuits in California state court or in administrative proceedings. The right of action is limited to injunctive relief and attorney's fees.

Takeaways

— Companies should monitor their use of compute carefully over time, as additional training runs or subsequent modifications can bring otherwise exempt AI models within scope.

— Companies should consider published frontier AI frameworks as imposing regulatory requirements given that TFAIA establishes liability for failing to adhere to the frameworks' terms.

— Companies should implement clear incident reporting policies and procedures to ensure they meet TFAIA's short reporting deadlines for critical safety incidents. These protocols should account for the possibility that OES ingests critical incident reports not only from frontier developers, but also from the public.

— Companies should consider reassessing catastrophic risk with each new potential use case or integration, since new deployments in sectors like transportation, healthcare, defense and critical infrastructure may trigger TFAIA's disclosure and reporting requirements.

— Updates to human resources and employment policies and practices may be required to integrate TFAIA's whistleblower protections.

If you have any questions regarding the matters covered in this publication, please reach out to any of the lawyers listed below or your usual Davis Polk contact.

Matthew J. Bacal
+1 212 450 4790
matthew.bacal@davispolk.com

David I. Feinstein
+1 212 450 3293
david.feinstein@davispolk.com

James W. Haldin
+1 212 450 4059
james.haldin@davispolk.com

David Lisson
+1 650 752 2013
david.lisson@davispolk.com

Howard Shelanski
+1 202 962 7060
howard.shelanski@davispolk.com
