October 14, 2025

California Enacts the Transparency in Frontier Artificial Intelligence Act (SB 53)

On September 29, 2025, Governor Gavin Newsom signed into law the Transparency in Frontier Artificial Intelligence Act (TFAIA), making California the first state to require public, standardized safety disclosures from developers of advanced "frontier" artificial intelligence (AI) models. In the absence of comprehensive federal legislation, California joins Colorado and Texas in advancing state-level AI governance. Unlike broader proposals considered in 2024, which included controversial provisions like mandatory third-party audits and “kill-switch” requirements, SB 53 focuses on transparency and accountability for developers of the most capable general-purpose foundation models and on harms that rise to the level of "catastrophic risk." Most requirements take effect January 1, 2026.

Why does TFAIA matter?

TFAIA requires public disclosures before deployment, timely reporting to the State when serious safety issues arise, and strong internal governance and whistleblower protections at qualifying developers. The statute is calibrated around a compute threshold that closely tracks emerging federal and international lines, signaling a de facto baseline for U.S. safety transparency. Companies that train or substantially modify large-scale models, or that partner with such developers, should treat SB 53 as a near-term compliance priority and align disclosures across related California measures.

Who does TFAIA apply to?

TFAIA applies to developers of “frontier AI models” which it defines as foundation models trained using more than **1026** integer or floating-point operations (“FLOPs”), including compute used for subsequent fine-tuning, reinforcement learning, or other material modifications. As of 2025, independent analysts estimate that only a small number of models have reached or exceeded the **1026** FLOP threshold total. One forecasting analysis from EPOCH AI projects around **~10** models at or above the **1026** threshold by 2026, growing rapidly thereafter as training budgets and clusters scale. The law imposes additional transparency and accountability requirements on “large frontier developers” whose annual revenue, including affiliates, exceeded $500,000,000 in the preceding calendar year. Coverage is not limited to California-based entities if models are made available to users in California.

What do frontier developers need to do?

Frontier AI framework (large frontier developers). Large frontier developers must publish, maintain, and follow a written **frontier AI framework** that explains how the organization incorporates national and international standards and industry-consensus practices; sets and evaluates thresholds for capabilities that could pose catastrophic risk; applies mitigations and tests their effectiveness (including through third-party evaluations); and assigns governance and decision-making responsibilities. The framework must also describe cybersecurity measures to secure unreleased model weights, criteria and triggers for updates, and

the process for identifying and responding to critical safety incidents. The framework must be reviewed at least annually, and any material modification must be posted within 30 days together with a brief justification.

Deployment-time transparency reports (all frontier developers). Before or concurrently with deploying a new frontier model—or a substantially modified version—developers must publish a **transparency report** (which may be presented as a system or model card) that identifies the model's release date, supported languages and modalities, intended uses, general use restrictions or conditions, and provides the developer's website and a contact mechanism. Large frontier developers must also summarize their catastrophic-risk assessments, disclose results, describe the role of any third-party evaluators, and explain other steps taken under the framework.



Critical safety incident reporting (all frontier developers). Developers must report qualifying **critical safety incidents** to the California Office of Emergency Services (OES) **within 15 days of discovery**. If an incident presents an imminent risk of death or serious physical injury, the developer must report within 24 hours to an appropriate public-safety authority. OES will maintain both public and confidential channels for submissions and will publish anonymized annual summaries beginning **January 1, 2027**. Incident reports and internal risk-assessment summaries are exempt from the California Public Records Act.

Internal-use risk summaries (large frontier developers). Large frontier developers must transmit periodic summaries to OES regarding assessments of **catastrophic risk from internal use** of frontier models (quarterly by default or on another reasonable written schedule).

Truthfulness and narrow redactions. The statute prohibits materially false or misleading statements about catastrophic risk and, for large frontier developers, about implementation of or compliance with the frontier AI framework. Public disclosures may be redacted only as necessary to protect trade secrets, cybersecurity, public safety, or U.S. national security, or to comply with law; developers must describe the character and justification of any redactions and retain unredacted materials for five years. (Good-faith statements that were reasonable under the circumstances are not violations.)

Whistleblower protections and internal channels. The law also adds new Labor Code provisions prohibiting retaliation and the use of gag clauses against **covered employees** who raise catastrophic-risk concerns or SB 53 violations with the Attorney General, federal authorities, or appropriate internal recipients. Large frontier developers must provide an anonymous reporting channel with monthly status updates to the reporter and quarterly briefings to officers or directors (with a carve-out when an officer or director is accused). Courts may award attorney's fees to successful plaintiffs, and injunctive relief is available and not stayed on appeal.

Enforcement and penalties. Only the California Attorney General may bring civil actions. Penalties may reach **$1,000,000 per violation**, scaled by severity. Loss of equity value does not constitute property damage for penalty purposes.

State infrastructure and local preemption. SB 53 establishes a consortium to design **CalCompute**, a state-backed public cloud cluster intended to support safe, equitable, and sustainable AI research. A report to the Legislature is due **January 1, 2027**, subject to budget appropriation. The statute also preempts local measures adopted on or after **January 1, 2025** that specifically regulate frontier developers' management of catastrophic risk.

When does this take effect?

Core publication, reporting, truthfulness, and whistleblower obligations apply beginning **January 1, 2026**. OES and Attorney General anonymized reporting and annual reviews begin **January 1, 2027**.

Other California AI regulations taking effect on January 1, 2026

AB 2013 (training-data transparency). Developers of generative AI systems made available to Californians must post training-data documentation on the developer's website for systems released on or after January 1, 2022. Compliance is required by **January 1, 2026**, and again upon any substantial modification.

SB 942 (AI content transparency). Covered generative-AI providers must implement content-transparency measures, including making available a free, publicly accessible detection tool that allows users to assess whether audio, image, or video content was created or altered by the provider's system. The law also contemplates durable provenance signals (for example, metadata) to support downstream disclosure. SB 942 takes effect **January 1, 2026**.

Pending California legislative proposals

AB 412 (copyright documentation for training data) would require developers to document copyrighted materials they knowingly use in training and to provide a mechanism for copyright holders to verify whether their work appears in training datasets.

SB 243 (companion chatbots) has passed both houses and awaits the Governor's action. It would require clear disclosures that users are interacting with AI, periodic reminders, and safeguards designed to restrict minors' access to sexual content.

Bottom Line

Frontier developers should promptly determine coverage by reconstructing training and post-training compute (including fine-tuning and RLHF) and by confirming consolidated revenue status. If any model meets the **1026 FLOPs threshold**, the frontier developer should: (1) draft, adopt, and publish a standards-aligned frontier AI framework that sets capability thresholds, mitigation gates, third-party evaluation criteria, and model-weight security controls; (2) operationalize a deployment-gating transparency workflow and a critical-incident playbook that satisfy the 15-day and 24-hour reporting clocks; and (3) implement whistleblower channels, employee notices, and non-retaliation safeguards consistent with the new Labor Code provisions. TFAIA's focus on catastrophic-risk management and alignment with federal compute thresholds positions it to become the de facto baseline for AI safety transparency in the United States.