Where Whistleblower Protections Fit Within California's TFAIA

High-powered artificial intelligence (AI) large language models (LLMs), also known as frontier models, are increasingly integrated into commercial, defense and critical infrastructure systems. Regulators are intensifying efforts to ensure responsible development and identify risks early. Beginning Jan. 1, 2026, California will implement the Transparency in Frontier Artificial Intelligence Act (TFAIA) (SB 53), the state's first statute aimed at governing high-risk AI models and the organizations that build and deploy them.

The law introduces new reporting, disclosure and safety obligations for frontier developers, reflecting California's expanding role as a testbed for AI governance. Unlike stalled federal proposals, TFAIA establishes state-driven expectations for internal risk reporting and transparency. The law aims to create guardrails against material catastrophic risks and critical safety incidents that could arise from the development and use of LLMs.

A key component of TFAIA is its whistleblower protections, which prevent an employer from taking an adverse action against a covered employee for engaging in a protected activity, such as reporting illegal activities or safety violations or assisting in an investigation. These protections empower employees to report AI safety and security risks without fear of retaliation, making early compliance planning essential for attorneys advising frontier developers.

What Is the Transparency in Frontier Artificial Transparency Act?

The TFAIA applies to two categories of companies developing advanced LLMs:

- **Frontier developers**: entities that train or initiate training of an advanced AI system
- **Large frontier developers**: entities that train or initiate training of at least one high-capacity AI system that meets high-compute thresholds and generate more than $500 million in annual revenue

The distinction determines which organizations face the statute's most stringent reporting and disclosure obligations. Large frontier developers are subject to mandatory reporting channels, employee training and formalized procedures to comply with whistleblower protections, while frontier developers must maintain transparency and incident reporting obligations without the same breadth of internal protocols.

Under the TFAIA, both types of developers must publicly disclose their safety and security protocols on their websites. These disclosures will need to include details about safety evaluations, risk-mitigation practices and security measures governing the model's development and deployment. If a developer discovers a "critical safety incident," they must submit a report to the California attorney general within 15 days of discovery. These critical safety incidents include:

- Death or serious injury to 50 or more people
- Damage exceeding $1 billion from a single incident
- Catastrophic system failures
- Loss of model control
- Cyberattacks that compromise safety mechanisms
- Situations where an AI model materially increases the risk of chemical, biological, radiological or nuclear weapon development

These heightened incident-reporting requirements create a compliance environment where employees are integral to identifying and escalating risks, reinforcing the role of whistleblower protections in mitigating catastrophic outcomes.

Whistleblower Protections Engrained in the Law

To encourage early reporting of critical incidents, the TFAIA includes explicit anti-retaliation protections for employees of frontier and large frontier developers who report safety concerns, risks or violations.

Under the statute, employees may not be retaliated against for:

- Reporting potential or actual catastrophic risks associated with a frontier AI model
- Disclosing internal safety failures
- Raising concerns about inadequate security protocols
- Providing information about the misuse of a model to create or assist in developing dangerous weapons
- Reporting loss of control, emergent capabilities or other failures that could create large-scale harm

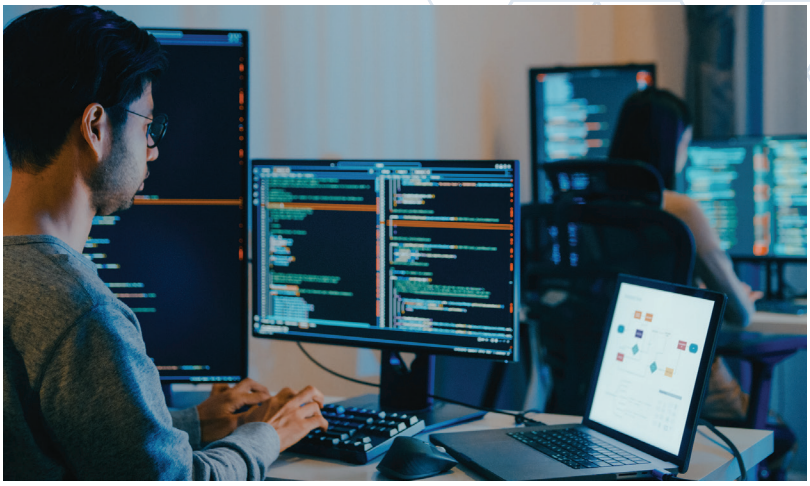
Employees of AI developers can report a system risk that violates the TFAIA externally to the California attorney general and federal authorities, or internally through supervisors and other established internal reporting channels.

These protections build on traditional whistleblower doctrines but are tailored to the unique risks AI systems pose, including national security, cybersecurity and catastrophic biological or nuclear harm. Retaliation, such as firing, demoting, cutting pay/hours, harassing or reassigning to undesirable roles, can expose developers to familiar civil and administrative consequences under California's whistleblower framework, including those in Labor Code section 1102.5, and may influence how regulators evaluate a company's safety controls, internal investigations and public disclosures.

Because the statute relies heavily on transparent reporting, enforcement agencies are expected to treat retaliation as a serious obstruction to the law's intent, increasing the risk of statutory penalties or reputational damage for developers.

Implications for AI Developers

To comply with the TFAIA's whistleblower provisions, frontier developers must implement internal systems to identify, escalate and respond to risk reports if they do not already exist.



Large frontier developers must maintain anonymous internal reporting processes that include clear communication channels for employees to report safety concerns. Channels should be accessible, confidential and monitored by trained personnel. Employees involved in AI development should receive training on catastrophic-risk indicators and safety-incident criteria.

For example:

- An engineer notices that an AI system controlling a regional power grid is generating outputs that could overload transformers, risking widespread outages and damages exceeding $1 billion. When the engineer reports this internally, the TFAIA protects them from retaliation while the company investigates and reports the critical safety incident.
- A developer identifies that an AI system exhibits emergent behavior, bypassing safety protocols and potentially taking harmful actions without human oversight. Internal reporting allows the company to assess and remediate the issue, with whistleblower protections shielding the employee from retaliation.
- A security analyst discovers that a deployed AI model is being exploited to launch automated cyberattacks affecting critical infrastructure. By reporting the incident, they are protected under the TFAIA while the company notifies the attorney general and takes corrective action.

Detailed logs will help companies demonstrate compliance and mitigate liability. Developers should document complaints, investigate, adopt remedial measures and determine whether the incident requires mandatory reporting. Documentation is critical if a retaliation claim arises.

Given the compliance burden, developers should ensure that HR, compliance and technical teams are well-trained to identify reportable AI-related risks, escalate concerns, protect employees from retaliation and fulfill statutory obligations to the attorney general.

Best Practices for Advising Clients

Attorneys should work closely with AI developers to assess current processes or establish a clear framework for evaluating and reporting incidents, including defining what is a reportable event, integrating whistleblower protections and mapping escalation and documentation processes. Policies should outline employee rights, definitions of reportable risks, internal complaint processes, assurances of nonretaliation, confidentiality and investigative procedures. Counsel should also advise on privilege considerations when preserving investigative materials.

Once an incident occurs, companies must compile supporting documentation efficiently. Attorneys can design templates and advise on preserving investigative materials and mapping when counsel must be involved for statutory or whistleblower matters.

Clients must maintain accurate, up-to-date public disclosures; misalignments could increase regulatory or litigation risk. They must also conduct evaluations and maintain privilege where appropriate.

Proactive planning reduces the likelihood of retaliatory behavior and the risk of enforcement actions. Attorneys can help clients anticipate incidents, document internal evaluations and maintain privilege where appropriate.

Whistleblower Protections Mark a New Era in AI Oversight

The TFAIA is more than California's first attempt to regulate frontier AI systems. It establishes a new model for transparency by empowering employees to report risks without fear of retaliation. For developers, protections shift how AI safety and compliance are managed. For attorneys, the TFAIA introduces new advisory responsibilities, from integrating processes and preparing AI developers for regulatory scrutiny to advising AI whistleblowers on reporting requirements and their rights under the law against employer retaliation.

As California continues shaping national AI regulation, the TFAIA's whistleblower provisions may serve as a model for future laws, making early compliance planning critical.