Published on *The National Law Review* https://natlawreview.com

California Passes Leading AI Safety Bill, Awaits Governor Approval

Article By:

Mallory Acheson, CIPM

Franklin Chou

Jason I. Epstein

The Safe and Secure Innovation for Frontier Artificial Intelligence Models Act ("AI Safety Bill") was passed by the California legislature on Aug. 28. Governor Gavin Newsom has until the end of this month to sign the bill into law. The AI Safety Bill introduces significant compliance obligations for businesses developing, training, or fine-tuning artificial intelligence models that meet specific computational power and financial thresholds and create critical harms, all as further defined, below.

Affected businesses should consider several immediate steps to ensure they remain compliant:

1. Assess your AI models for risks of causing critical harm (defined below).
2. Develop and implement safety protocols that address cybersecurity risks and maintain oversight over AI model behavior.
3. Plan for pre-deployment and post-deployment compliance, including annual audits, reporting requirements, and risk assessments.
4. Prepare for enforcement actions: non-compliance could lead to significant penalties, including fines as a percentage of computing power costs.

In the detailed analysis below, we'll explore the key provisions of the AI Safety Bill and how they shape the responsibilities of businesses working with covered models.

The AI Safety Bill establishes a Board of Frontier Models (the "Board"), a component of the Government Operations Agency, which will provide oversight and regulation over individuals or entities that develop, train, or fine-tune covered models. Under the bill, "covered models" refers to AI models that meet the following criteria:

- Before Jan. 1, 2027:
  - Any AI model trained with computing power exceeding 10^{26} integer or floating point operations, with costs exceeding $100 million; or
  - Any AI model fine-tuned by adjusting the weights of an existing model, requiring

computing power exceeding 3 x 10^{25} integer or floating-point operations, with costs exceeding $10 million.
- After Jan. 1, 2027, the cost thresholds will remain the same (with adjustments for inflation) and the compute thresholds will be as the Board may determine (or remain unchanged if no threshold is set by the Board).

For purposes of this article, “covered model” includes the covered model and any derivative of the covered model.

The AI Safety Bill is directed at preventing “critical harm,” which is defined as:

- The creation or use of weapons of mass destruction (chemical, biological, radiological, or nuclear), leading to mass casualties.
- Cyberattacks on critical infrastructure resulting in mass casualties or damage of at least $500,000,000 (with adjustments for inflation), where the covered model either conducts or provides instructions for the attack.
- Any covered model acting with limited human oversight, causing mass casualties or property damage over $500,000,000 (with adjustments for inflation) that would qualify as a crime if committed by a human.
- Other severe harms to public safety comparable to the above.

In addition, the AI Safety Bill defines “AI safety incident” as any incident that demonstrably increases the risk of critical harm by any of the following means:

- The covered model autonomously engaging in behavior other than at the user’s request; or
- Any theft, misappropriation, malicious use, inadvertent release, unauthorized access, or escape of the model weights of a covered model; or
- The critical failure of technical or administrative controls pertaining to the covered model, including controls limiting the ability to modify the covered model; or
- Unauthorized use of the covered model that would cause or materially enable critical harm.

Key Requirements for California Companies Developing Covered Model

For California companies developing, training, or fine-tuning covered models, these are some of the compliance obligations under the AI Safety Bill and where they relate to the AI lifecycle:

- **Pre-Training Compliance:** Before training a covered model:
    - Assess whether the covered model is reasonably capable of causing or materially enabling a critical harm; if there is an unreasonable risk that the covered model might cause or materially enable a critical harm, such model cannot be used for any commercial or public use; and
    - Implement reasonable administrative, technical, and physical cybersecurity protections to prevent unauthorized access to, misuse of, or unsafe post-training modifications of, the covered model; and
    - Develop the capability to immediately shutdown and cease operations of the covered model; and
    - Develop and implement a written protocol (the “safety and security protocol”) that will mitigate the risks of any critical harm posed by the covered model and disclose such protocol to the Attorney General; and
    - Undertake all other reasonably appropriate measures to prevent the covered models

    from posing unreasonable risks of causing or materially enabling critical harms.

- **Pre-Deployment Compliance:** Before the covered model is deployed or used in any commercial, public, or foreseeably public use:
    - Take reasonable steps to prevent the covered model from causing or contributing to critical harm; and
    - Take reasonable steps to accurately attribute actions (including any critical harm that may result therefrom) to the covered model; and
    - Conduct annual third-party audits to ensure the safety and security protocol and the covered model to which such protocol applies, continues to adequately evolve in the context of advancements in science, technology, and national and international standards in accordance with regulations which may be further promulgated; and
    - Disclose training data information.
- **Post-Deployment and Ongoing Compliance:** After the covered model is deployed:
    - Within 30 days of initial deployment and annually thereafter, submit a statement of compliance with the requirements of the AI Safety Bill; and
    - Continuously assess the risk of critical harm posed by the AI model; and
    - Report to the Attorney General any AI safety incident within 72 hours; and
    - Retain a copy of the assessment analyzing the extent to which the covered model is reasonably capable of causing or significantly enabling critical harm for the entire period the model is available for commercial, public, or foreseeable public use, plus an additional five years; and
    - Retain a copy of the safety and security protocol for the entire period the model is available for commercial, public, or foreseeable public use, plus an additional five years; and
    - Review and update the safety and security protocol at least annually to ensure the provisions are right sized to prevent critical harm as the AI model evolves.

Enforcement and Compliance Expectations

The California Attorney General is empowered to seek a range of remedies for violations. These include civil penalties for violations causing death, bodily harm, or significant property damage, with fines reaching up to 10% of the computing power costs used to train the AI model, increasing to 30% for subsequent violations. Additional penalties include fines for labor violations, with amounts up to $10 million for related offenses. The Attorney General may also pursue injunctive relief, monetary damages, punitive damages, and attorneys’ fees. Courts are authorized to provide any other relief deemed appropriate and may disregard corporate formalities to impose joint and several liability if affected businesses have structured their entities in a way that unreasonably limits or avoids liability.

The AI Safety Bill introduces new regulatory requirements for large-scale AI models developed or fine-tuned in California. It establishes compliance obligations and grants the Attorney General the authority to bring civil enforcement actions for violations. Affected California businesses need to understand how these regulations will impact their operations, including the need for robust safety protocols and potential penalties for non-compliance. The interaction between these state-level rules and existing or future federal regulations is still uncertain and companies should closely monitor how this regulatory framework evolves.

Companies must proactively prepare for compliance. This involves not only understanding and adhering to the new regulations but also enhancing internal safety protocols. By anticipating these regulatory changes and integrating robust risk management strategies, businesses can better navigate the evolving landscape of AI regulation and safeguard their operations against potential liabilities.

National Law Review, Volume XIV, Number 268

Source URL:https://natlawreview.com/article/california-passes-leading-ai-safety-bill-awaits-governor-approval