SB 1047: What You Need to Know

SB 1047, the Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, is a bill currently progressing through the California state legislature. Proposed by State Sen. Scott Wiener, the bill aims to support responsible AI innovation in the state while addressing critical risks. This document offers a concise overview of what the amended bill proposes.

Q. Which systems are covered?

AI models that would be covered under this legislation are limited to those that are trained with >$100 million and >10^26 integer or floating-point operations in computing power. Models that are fine-tuned using over three times 10^25 additional operations at a cost of over $10 million would also be covered as new models - this is mostly relevant to open source models being fine-tuned with significant computing power. There are currently no AI models available that would reach the threshold the bill outlines.

→ Beginning in 2027, the Government Operations Agency will annually re-evaluate these thresholds for covered models, to keep up with the state of science. However, there will always be a $100 million monetary threshold, and a $10 million monetary threshold for fine-tuning, with adjustments for inflation. A Board of Frontier Models, made up of independent experts appointed by the Governor and state legislature, will advise the Government Operations Agency on appropriate thresholds.

Q. What exactly are the safety requirements for developers?

The bill requires developers to write and follow a safety and security protocol before training a model that would be covered, outlining how they will evaluate for and manage risk of critical harm that the model may present. The protocol must be submitted to the Attorney General's office, and meet certain specifications, but companies get to write their own protocols.

→ Pre-deployment safety evaluations, with much discretion given to the AI developer, must be part of this protocol.

Q. What constitutes a "critical harm"?
The bill defines critical harm as harm from a covered model creating or enabling cyberattacks, CBRN weapons, or its own autonomous actions that results in mass casualties and/or at least $500 million in damages.

→ Importantly, it **does not include** harms caused by AI outputs if the information is already publicly available, nor harms not directly related to the developer's actions in creating, using, or releasing the AI model.

Q. What is “shutdown capability"?
Developers would be required to incorporate an “emergency shutdown” capability when training a covered model.

→ Importantly, this only applies to models that remain under a developer’s control, which **exempts** most **open-source models** from this provision.

Q. Who will ensure compliance?
Third-party auditors will be responsible for evaluating developers’ compliance with the safety and security protocol they’ve written, and will submit reports on compliance to the Attorney General. Auditors themselves will be held to standards for practice published by the Government Operations Agency. The Board of Frontier Models will advise the Government Operations Agency on appropriate regulations and standards for auditors.

Q. What are the reporting requirements?
The bill would require AI developers to report AI safety incidents (i.e., an incident which “demonstrably increases the risk of a critical harm occurring”) within 72 hours.

Q. What is CalCompute?
The bill would establish a consortium to develop a public cloud computing cluster, CalCompute, to offer computing power and resources for AI innovation and research, particularly for the public sector and smaller companies that may not be able to afford the significant resources required to train AI models.

Q. What does it do for whistleblowers?
The bill would introduce whistleblower protections for workers at organizations developing AI models. Workers would be allowed to report both violations of the proposed legislation *and* risky behaviors/actions that might not constitute a violation, to the Attorney General or the Labor Commissioner, and could not face retaliation for it.

Q. What penalties could developers face for violations?
Developers would not face any civil penalties for violations of SB 1047 in the absence of actual harm or an imminent risk or threat to public safety. Additionally, the bill no longer includes any criminal violations, only civil, and only some of those have penalties associated.

➔ A court can, however, order injunctive relief before a critical harm occurs to mitigate unreasonable risk (e.g., demanding that a developer rectify an egregious vulnerability, or mitigate an unreasonable risk of critical harm before continuing commercial distribution of the covered model).

Further reading:
Bill SB 1047, updated 8/19/2024.
What does SB 1047 do and why is it needed?, SafeSecureAI.
Guide to SB 1047, Zvi Mowshowitz.