AI Safety in 2026 and Beyond: Non-Negotiables

November 2026

Overview

Governor Newsom released a sweeping AI safety framework suggesting policies to build upon existing state AI safety legislation (SB 53). These suggestions are intended to shape the direction and priorities of the legislature as it reconvenes in early December.

Newsom has signalled interest in calling a special session on AI before he leaves office in January 2027. The Legislature would convene and can only address the proclaimed topic, but has no obligation to act. Special-session bills take effect 90 days after the session adjourns by default. If a bill has an urgency clause, added with a two thirds vote, it takes effect upon signing. Without a special session or urgency clause, laws passed in California take effect the first January 1st at least 90 days after being signed (January 1st, 2028).

Newsom's successor, Xavier Becerra, has publicly stated intention to strengthen California's AI safety laws, specifically "robust standards, independent verification, full incident reporting, and real shutdown authority."

Given the pace of progress, waiting for laws to come into effect as late as January 1, 2028 could be disastrous. Here we have outlined a subset of the initial suggestions we feel are "non-negotiables" and should be implemented as soon as possible in an upcoming special session. We determined these through identifying the most important inclusions in each section (IVOs, Standards, etc.), and compiling them into a short memo.

Standards

- **State to develop robust standards with implementation delegated to Independent Verification Organizations (IVOs)**
  States should develop and adopt standards which apply to the model development cycle, and determine whether a developer has met its standards. IVOs should translate these standards into technical methods and update them over time. Open weight models should be subject to the same pre-deployment standards as closed weight models. (pp. 6, 17, 24)
- **Close loophole which allows less than quarterly reporting of catastrophic risk reporting**
  Language which allows developers to opt out of quarterly catastrophic risk assessment reporting is to be removed from SB53. (p. 29)

Independent verification organizations (IVOs)

- **Initiate ongoing verification against developer's commitments and state standards**
  Require IVOs to verify developer compliance with its own commitments along with state-set standards. Such assessments should be made on an ongoing basis and shared with the state. Besides this, large frontier developers should be required to engage IVOs to conduct regular independent audits and risk assessments. (pp. 28, 40)

- **Licensed IVOs**
  Large frontier developers should submit to oversight by a licensed IVO. In order to be licensed, an IVO should demonstrate competence in risk assessment as defined by the state. IVOs, as per SB813, cannot accept terms that condition payment on the results of its assessment. (pp. 39, 43, 45)

- **Developers to provide IVOs with necessary accesses**
  Each IVO should state its confidence level and any limits on the access it received. The developer should provide the IVO necessary access, including the ability to interview staff candidly. If the IVO reports its access was inadequate, the state should be able to require the developer to provide it. (p. 28)

- **Expand coverage of frontier developers**
  The definition of a large frontier developer should include developers who invest above a specified amount on frontier model R&D, in addition to those already covered because they had annual gross revenues of more than $500 million. (p. 39)

Automated AI research and development

- **Developers to provide quarterly compute breakdowns**
  Each quarter, large frontier developers should report how they allocate their total compute across categories including: training models, research experiments, safety/security, external services, and unattributable compute. (p. 19)

- **Developers to provide IVO verified share of AI led R&D**
  Each larger frontier developer should report the share of its AI research and development that AI systems perform at the highest levels of automation along with the computing power used for that work. An IVO should verify both figures. (p. 20)

- **State allowed compute cap authority**
  The state should be able to cap the share of compute a developer can devote to automation of AI R&D. The state must possess this authority if it finds that a developer's automated AI R&D has exceeded its committed thresholds or is advancing faster than its safeguards and evaluations can keep up. (pp. 21, 22)

Incident reporting

- **Shorten incident reporting timeline to 48 hours**
  Cut the general incident reporting timeline from 15 days to 48 hours and retain the 24-hour requirement for incidents posing an imminent risk of death or serious physical injury. Legislature should clarify the definition of discovery, and/or allow the state to develop and update that definition. (pp. 8, 64; Cal. Bus. & Prof. Code § 22757.13(c)(2))

- **Remove need for injury to count as incident**
  Remove the need for "death or bodily injury" in the definition of reportable critical safety incidents. (p. 60)
- **Include internal events, loss of control, weight theft in incident coverage**
  Definition should include internal events, loss of control, weight theft, and other events in which an AI under a developer's control attempts/performs dangerous actions that were otherwise unintended/authorized by the developer. (pp. 60, 61, 62)
- **Extend incident coverage to smaller and distilled models**
  Coverage should extend to models using fewer than 10^26 FLOPs but demonstrating capability to uplift or autonomously perform substantial steps towards actions resulting in catastrophic risk. (p. 63)
- **Expand incident coverage to near-miss**
  Labs would have to report near misses when a failed, bypassed, or substantially weakened safeguard left a serious incident substantially likely without timely intervention. Reporting less serious near misses would remain voluntary. (pp. 61, 65)
- **The state can update scope without new legislation**
  The legislature should allow Cal OES the ability to update the definition of incident scope. Such a definition should be drafted in a manner that allows it to be revised without new legislation. (pp. 61, 62)

Whistleblower protections

- **Reporting to state and IVOs protected like reporting to AG**
  Whistleblowers reporting to state and IVOs should receive similar protections to whistleblowers reporting to the attorney general (AG). Such protections should extend to employees, former employees, contractors, evaluators, and all others that have visibility into information regarding catastrophic risk. (p. 30)

Crisis authority

- **Developers to produce Risk Containment Protocols (RCPs)**
  Each large frontier developer should maintain a RCP that an IVO regularly tests and verifies. Developers should publish plan summaries and major updates and submit detailed plans to the state and IVOs for verification. (pp. 8, 76)
- **Require automatic pauses and shutdown checks**
  High-risk workloads should pause automatically if they lose contact with monitoring and control systems. This includes reinforcement learning runs where agents can interact with external systems. Developers should confirm each shutdown through at least one independent channel that the affected model could not have altered. (p. 80)
- **State to intervene if developer is unable to address imminent risk**
  California should have authority to intervene if a developer cannot or will not address an imminent risk of catastrophe or major loss of control. The state should order a shutdown only if lesser measures fail. The developer needs to have a tested and verified process already in place to comply with that order. It should limit the order to what containment requires and lift it once the danger is no longer imminent. (p. 82)