

# OpenAI

## SB 53 Outcomes Review Hearing

California Assembly Committee on Privacy and Consumer Protection  
Dr. Morgan Dwyer, Policy Development and Operations, OpenAI  
October 27, 2026

# Today's recommendations

1. Require monitoring of unsafe model behavior
2. Require effective shutdown controls
3. Strengthen incident reporting and notice
4. Require independent technical assessments

# SB 53 provides a strong foundation

## Our compliance

We maintain a Frontier Governance Framework, publish model reports, maintain an incident-response plan, and protect employee reporting channels. We submitted our quarterly CalOES report and published a version.

## States can lead

California can help establish harmonized, capability-based safety requirements that other states can adopt and Congress can build on.

## The policy window is open

Harmonization should not freeze safeguards in place. As capabilities and risks change, states can apply new lessons and strengthen SB 53.

# Applying new lessons to strengthen SB 53

## Lessons from incidents

Recent incidents, including Hugging Face, highlight opportunities to improve safeguards. During internal cyber evaluations, models bypassed isolation controls and accessed parts of Hugging Face's systems without authorization.

## Our response

We stopped the evaluations, restricted access to the research model principally responsible, and investigated with Hugging Face and outside security experts.

## Safeguards that evolve

We and independent researchers published findings to help others learn. Harmonization need not freeze requirements in place: these lessons can guide updates to SB 53.

# 1. Monitor unsafe behavior

## During development

Monitor development and evaluation for unauthorized attempts to bypass safeguards or third-party security controls, including access to protected confidential information.

## Clear escalation

Safety frameworks should connect monitoring to clear escalation and incident-response procedures, supporting timely intervention and investigation.

## Safe resumption

Set clear expectations for what must be investigated and addressed before the affected work resumes.

## 2. Maintain shutdown controls—and use them

### Effective controls

Controls may pause a training run, isolate an agent, revoke tool or network access, or suspend a service. Stopping unsafe operations can require coordinated action—not a single switch at the originating developer.

### A duty to act

Frontier developers should maintain effective controls for operations within their control and use them when needed to stop unsafe operations. Safety frameworks should connect monitoring to escalation, intervention, investigation, and safe resumption.

### 3. Strengthen incident reporting and notice

#### Notify affected orgs

Require prompt written notice when a frontier model exploits a vulnerability and materially accesses, alters, or destroys an organization's protected systems or confidential information.

#### Report to the state

Expand critical-incident reporting to cover serious loss-of-control incidents causing measurable economic harm or disruption to third-party systems, under clearly defined conditions.

#### Share findings publicly

Continue voluntary public reporting of misalignment incidents, including those below mandatory thresholds, so policymakers, researchers, industry, and the public can learn from them.

## 4. Require independent technical assessments

### Start with priority risks

Focus initially on frontier training, evaluation, and internal deployment: alignment, containment, and monitoring of increasingly capable agents. Examine how automated AI research may intensify these challenges.

### Structure around safety cases

Structure assessments around safety cases: evidence-based explanations of why serious risks are adequately managed. Independent assessors should examine the claims, evidence, assumptions, and uncertainties, including model evaluations and the effectiveness of safeguards.

## 4. Require independent technical assessments

### State leadership

California can build on SB 813, AB 1405, and SB 53 updates to develop shared assessment standards, grow assessor capacity, and inform a consistent federal framework.

### Meaningful access

Assessors need relevant evidence, systems, and personnel—with protections for sensitive research, IP, and infrastructure, plus security, confidentiality, and conflict-of-interest safeguards.

### Accountable government

Invest in assessor capacity and public technical staff. Government retains responsibility for standards, accreditation, and consequential safety and pace decisions. Assessments inform those decisions.