Secure AI Project

How California Can Meet The Moment On Frontier AI Safety

Secure AI Project's legislative recommendations for California.

October 7, 2026 | Contact: info@secureaiproject.org

About Secure AI Project

> We develop and advocate for pragmatic policies to reduce risks of severe harm from advanced AI.



Thomas Woodside

Cofounder



Juan Carlos Salamanca

Advocacy



Andrew Doris

Advocacy

+16

Other employees



LEGISLATIVE TRACK RECORD	**California**	**New York**	**Illinois**
	SB 53, AB 1864. SB 1047 previously.	RAISE Act.	SB 315.

Secure AI Project

Our legislative framework

Scoping, verification, standards, orders, and pacing.

Scoping: Cover all frontier models, developers, and incidents

Recommendation 1

01

"Frontier model"

- SB 53 covers models trained with more than 10^{26} operations.
- The public has no evidence today's frontier models meet that bar.
- Adjust the definition to cover all frontier models, and allow it to be updated by regulation.

02

"Large frontier developer"

- Revenue is not well-correlated with risk.
- Well-funded labs like Safe Superintelligence Inc. have billions in investment but no revenue.
- Replace it with a $1B AI R&D spending threshold, which correlates far more strongly with catastrophic risk.

03

"Critical safety incident"

- No incident has ever clearly qualified as a "critical safety incident" under SB 53.
- We are left with only voluntary disclosure.
- Expand the definition to include more loss of control incidents.

Verification: Embedded IVOs

Recommendation 2

OVERVIEW

- Access to all relevant information to assess severe risks.
- Continuous assessment plus regular reports.

OUTPUTS

- **Annual audit** to assess if the developer complied with the law, including minimum safety standards.
- **Quarterly evaluation:** to assess if the developer's models pose serious risks, even if all laws have been followed.
- **At any time:** IVOs should always be able to report legal violations or imminent severe risks.

ASSIGNMENT

- Inherent incentive problems if companies choose their own IVOs.
- The state should assign the most qualified evaluator for each function.
- Companies pay for IVOs through a process managed by the state.

Minimum Safety and Security Standards

Recommendation 3

Why minimum standards?

- Transparency alone is no longer enough.
- Developers have weakened their commitments.
- A lack of minimum standards creates a race to the bottom on safety.

Rulemaking process

- OES adopts standards with input from IVOs, industry, and potentially nongovernmental interstate standard-setting bodies.
- IVOs verify compliance with standards.
- Given speed of advancements, statute should deem regulations to automatically qualify for emergency status where needed.

Inherently incomplete

In a field moving this quickly, it's impossible to do safety purely by checklist.

Orders and a “kill switch”

Recommendation 4

A WORKABLE “KILL SWITCH”

- A literal "switch" doesn't make sense for a number of reasons.
- We can and should require a documented shutdown procedure that would actually work in an emergency.

ORDERS

- A shutdown procedure isn't reliable if activation is entirely up to companies.
- We need a more flexible way than rulemaking to resolve disputes and prevent emergencies.

EXAMPLE TYPES OF ORDERS

- **Corrective action:** Require a developer to implement a corrective action or cease an activity creating risk.
- **Information access:** Require a developer to provide specified information to the IVO or to OES itself.
- **Removing redactions:** Require an IVO to remove a redaction made without good justification by the IVO or the developer.

Most orders take effect after 15 days to leave time to appeal. Emergency orders take effect immediately.

Pacing capabilities so safety can keep up

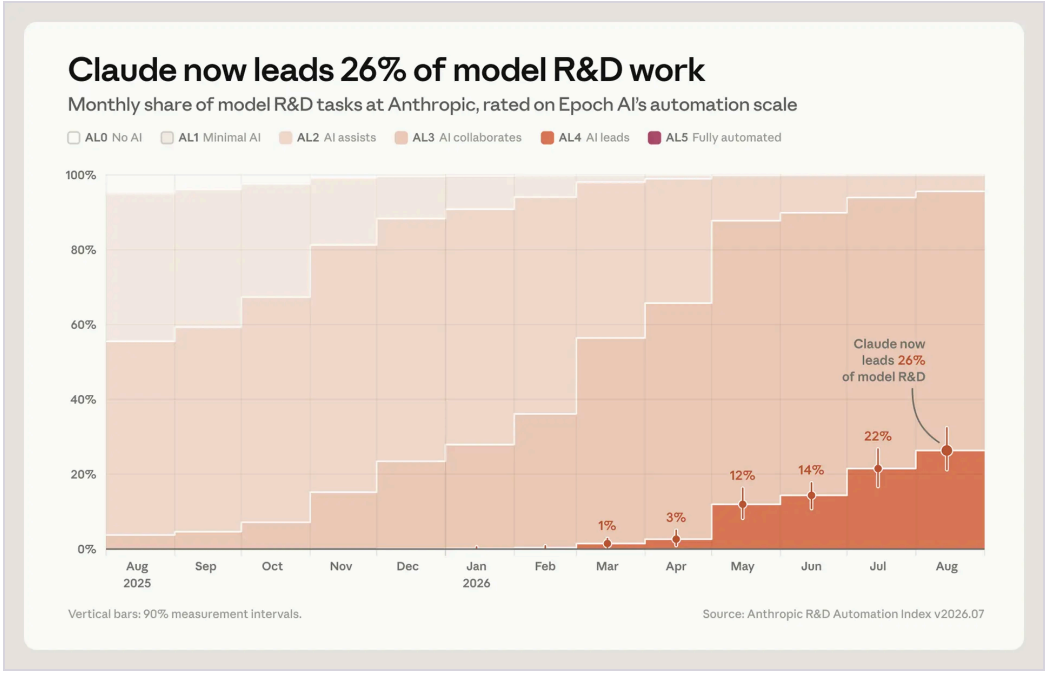
Recommendation 5

THE RISK

- Fully automated AI research could lead to recursive self-improvement.
- Without intervention, progress could be too fast to control.
- Employees at frontier AI companies have called for “pacing.”

POLICY STEPS

- **Measure** the speed of AI advancement.
- **Assess** a maximum safe rate of progress.
- **Limit** compute allocated to automated AI R&D to keep progress in safe limits.



Secure AI Project

We would love your support.

Questions?

secureaiproject.org | info@secureaiproject.org