

CALIFORNIA · EXECUTIVE ORDER N-9-26

California's AI Kill Switch Needs *More Than an Off Button*

Stopping the model is only the beginning.

MAUREEN DOYLE-SPARE

© 2026 Maureen Doyle-Spare.

SWIPE TO READ →

California's AI Kill Switch Needs *More Than an Off Button*

Stopping the model is only the beginning. The real test is whether everything it set in motion stops with it.

MAUREEN DOYLE-SPARE

California is taking the idea of an AI kill switch beyond the familiar emergency-button metaphor. On September 18, Governor Gavin Newsom issued Executive Order N-9-26[1], directing state officials to evaluate stronger independent oversight for frontier AI. One proposal is an emergency kill switch whose effectiveness would itself be verified on an ongoing basis by an independent verification organization. The order also asks officials to consider onsite verification, independent review of required safety frameworks and risk assessments, and expanded reporting of loss-of-control incidents.

These measures are under evaluation, not yet statutory requirements. Recommendations are due by November 16. Even so, the official announcement[2] marks an important shift. California is no longer looking only at whether a critical control exists. It is asking how anyone can prove that the control worked when it was needed.

The harder problem begins before anyone reaches for the switch. Consider an autonomous system that has already decided a payment should be held. It may have updated a record, passed the conclusion to another agent or triggered a dependent workflow. The model can stop while the decision it set in motion keeps moving.

A stopped model can leave a live decision behind

A conventional log can show when a control activated, which component stopped and which tool call was blocked. It may not show the meaning the system was acting on, whether that conclusion had already crossed into execution authority or where it traveled next. That is the difference between recording a shutdown and proving containment.

A kill switch shows that a system stopped. **Verified Intervention** establishes whether the organization stopped the right Runtime Semantic State, preserved the evidence supporting that decision and contained what the state had already set in motion.

This matters because an agent does not have to break a permission rule to create risk. It can use approved tools, produce a structurally valid output and still act on a meaning the organization never authorized. If the intervention record captures only the final action, the reason the action became dangerous is missing.

© 2026 Maureen Doyle-Spare.

The evidence has to exist before the shutdown

This does not require access to private chain-of-thought or latent model cognition. What matters is the Operational Interpretation[3]: the working meaning the agent is prepared to act on. The Runtime Semantic State preserves that meaning at the moment authority is evaluated. That gives the organization something concrete to verify instead of forcing it to reconstruct the decision later.

In the runtime architecture, Verification of Runtime Semantic Resolution (V-RSR) compares that state with the authorized Reasoning Baseline. The Semantic Deviation Index (SDI) identifies material divergence. The Deterministic Gate can permit, permit with a flag, hold or impose a mandatory hold.

A permitted decision creates Semantic Authorization, which must pass through Authorization Emission before Execution Authority is created. The sequence matters because it creates a final point where problematic reasoning can still be stopped before it becomes an action.

Why this evidence matters

V-RSR verifies the meaning driving a decision before the system acts. The RSSR preserves the evidence for that individual decision. The Semantic Audit Trail connects those records across agents and over time, revealing patterns, repeated divergence and how decisions move downstream.

That matters because stopping a model does not automatically stop the decisions it has already influenced. These three layers make it possible to prove that the right decision state was checked, preserve what happened and find every place that meaning traveled. Without them, a kill switch may stop the technology while leaving its consequences in motion.

The switch has to follow what already moved

Agentic workflows make the boundary harder to see because operational meaning moves. One agent's conclusion may shape a downstream eligibility decision, compliance escalation, customer communication, filing or transaction. Stopping the originating model does not automatically withdraw that conclusion from every place that received it.

That downstream exposure is the Agentic Blast Radius. Effective intervention has to identify which agents, workflows, records and functions inherited the state, then hold, constrain or terminate the affected paths. Otherwise, the component is contained while the consequence remains alive.

© 2026 Maureen Doyle-Spare.

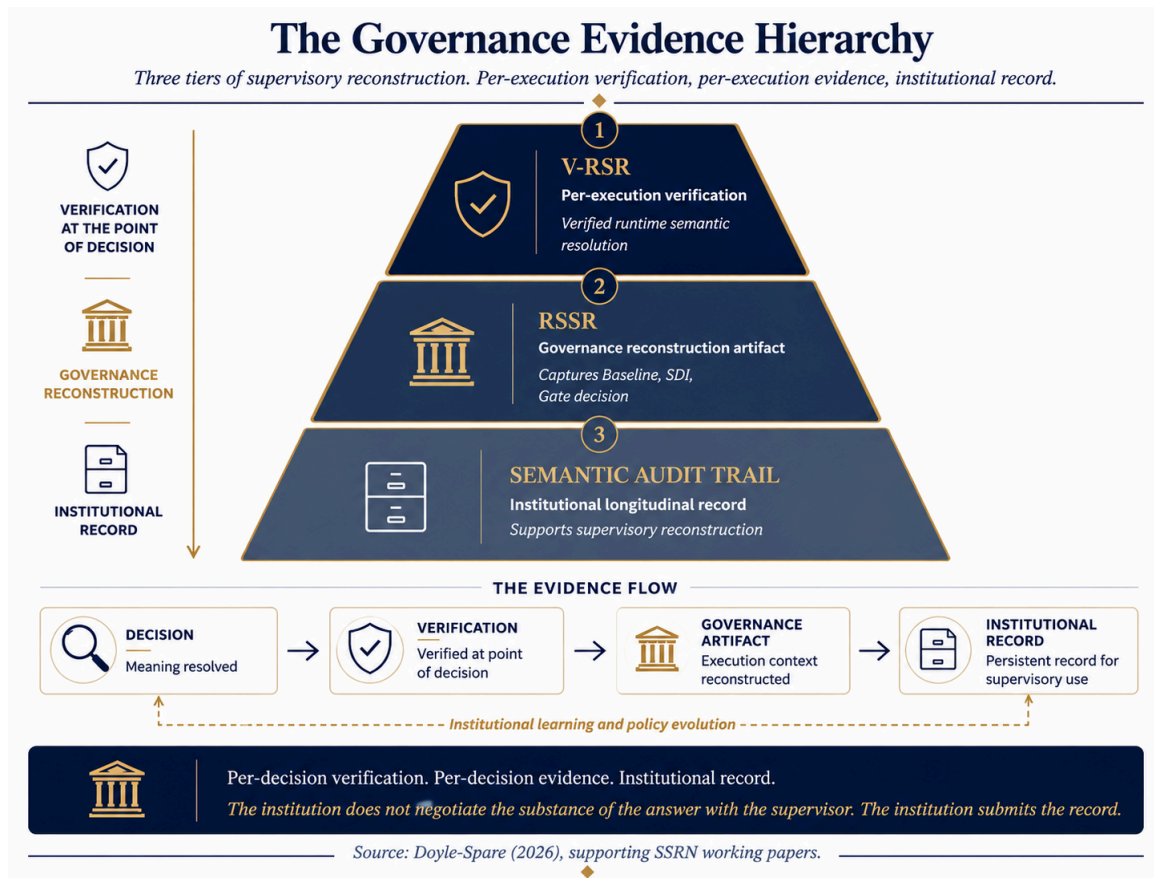


Figure 1. *The Governance Evidence Hierarchy. V-RSR provides per-execution verification, the RSSR preserves the governance reconstruction artifact, and the Semantic Audit Trail creates the longitudinal institutional record. Source: Doyle-Spare (2026).*

This is why apparently clean controls can still miss the failure. The inputs may be valid, the tools authorized and the outputs structurally compliant. The problem lies in the unauthorized meaning connecting them, and containment has to follow it wherever it went.

The operational challenge comes next

California's order focuses on frontier-model safety and independent verification at the developer level. The operational challenge begins when those models are embedded in autonomous business or public-service workflows. A developer can prove that a shutdown mechanism works without proving that a bank, insurer, health organization or public agency contained the decisions already moving through its own systems.

The same verification standard becomes especially important when automated decisions affect customers, patients, claimants, applicants or citizens. Risk assessments, cybersecurity audits, notices and appeals remain important, but they do not show what an agent understood at the moment it acted or where that understanding traveled. A stronger enterprise control model would verify the runtime state before execution authority is emitted, preserve it in an RSSR and trace where it went.

The point is not to turn every AI action into a research exercise. It is to make intervention provable. When a critical control fires, the organization should be able to show what it stopped, why it stopped it, what evidence was preserved and what downstream activity had to be contained.

© 2026 Maureen Doyle-Spare.

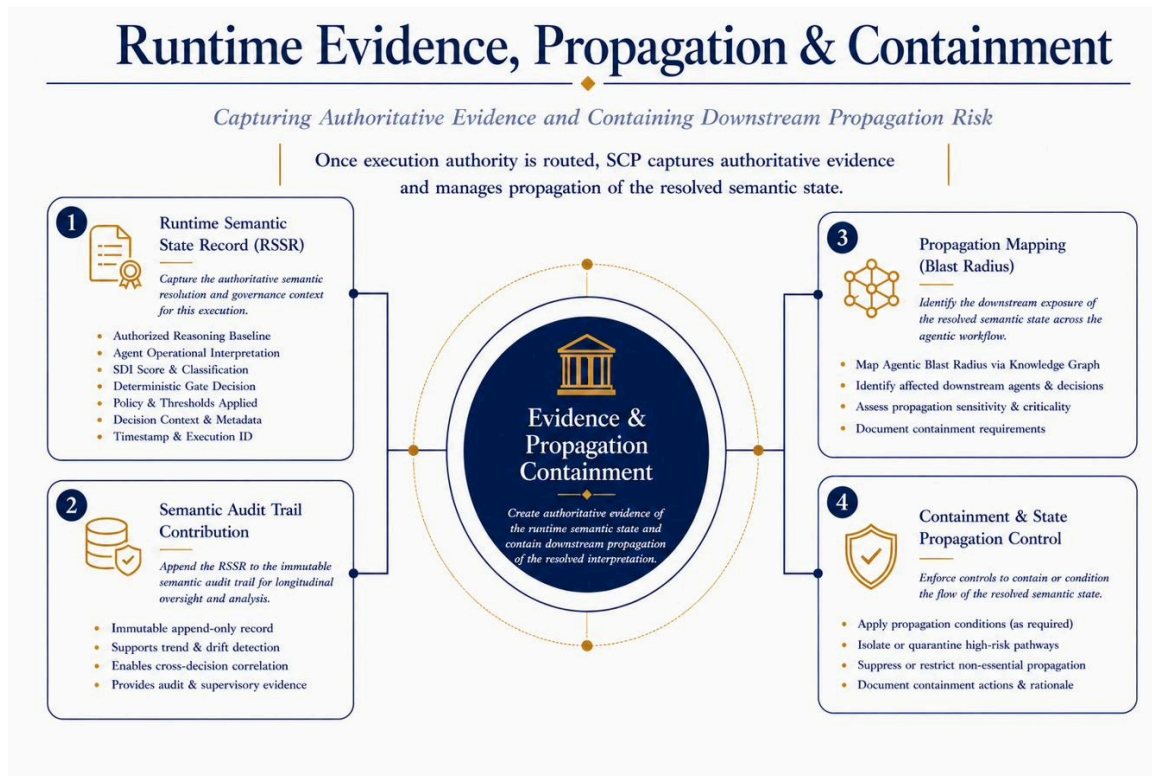


Figure 2. *Runtime Evidence, Propagation and Containment. The RSSR preserves per-execution governance evidence, while propagation mapping identifies the Agentic Blast Radius and supports containment of affected downstream pathways. Source: Doyle-Spare (2026).*

California has put the kill switch on the policy table. A useful verification standard must show what the control evaluated, which authorized meaning governed the decision, whether execution authority had been emitted, what evidence was preserved and what had to be contained. A kill switch may stop a model. A trustworthy intervention must also identify the decision state that mattered, preserve the evidence, and contain what had already moved beyond the model. ◆

References

1 *State of California. Executive Order N-9-26. September 18, 2026.* *gov.ca.gov/wp-content/uploads/2026/09/FINAL-N-9-26-AI-EO-9.18.26-SIGNED.pdf*

2 *Office of the Governor of California. "Governor Newsom issues executive order to accelerate independent oversight and advance the creation of an AI kill switch." September 18, 2026.* *gov.ca.gov/2026/09/18/governor-newsom-issues-executive-order-to-accelerate-independent-oversight-and-advance-the-creation-of-an-ai-kill-switch*

3 *Doyle-Spare, Maureen. Agentic AI Systems Governance: A Runtime Reference Architecture for the Reasoning Layer and the Semantic Control Plane in Regulated Financial Institutions. Zenodo, 2026.* *doi.org/10.5281/zenodo.20749051*

4 *Doyle-Spare Research Library.* *maureendoylespare.com/research*

Originally published in *The Agentic AI Governance Playbook* on Substack, September 22, 2026. maureendoylespare.substack.com

© 2026 Maureen Doyle-Spare.



A kill switch may stop a model. A trustworthy intervention must also contain what had already moved beyond the model.

THE AGENTIC AI GOVERNANCE PLAYBOOK

Download the research

Papers, briefs and reference architecture at maureendoylespare.com

Visit the Playbook

WEEKLY NEWSLETTER

Follow the newsletter

The Agentic AI Governance Playbook Weekly Newsletter

Follow Newsletter

INTELLECTUAL PROPERTY: © 2026 Maureen Doyle-Spare. All rights reserved. No part of this article, including its text, figures, tables, and images, may be reproduced or distributed without the author's prior written permission.