Safety Is Practical for All: A Verified Stop for Advanced AI

WP-VS-001 · Prepared for the Office of Rep. Ted Lieu by Sworne™ · pac@sworne.io · an Atom Works development center · Oct 8, 2026

Executive summary

A verified stop is the one AI safety measure every party can support today, because it asks nothing of anyone except proof that an off switch works. It does not regulate what a model may say, how it is trained, or who may build one. It asks that an advanced AI system can be stopped by a party it cannot influence, that the stop is rehearsed on a schedule, and that the rehearsal leaves a record someone outside the company can check.

California has already called for AI kill switches whose efficacy is verified on an ongoing basis by independent organizations. This paper proposes that the federal bill adopt the same definition, written as five plain outcomes, so a company that complies in California complies nationally and no preemption fight follows.

The five criteria: authority sits outside the system; the stop needs no cooperation; silence stops it; restart takes a second person; drills leave a record anyone can verify offline. Operators who meet them earn a verified-record defense; operators who do not are exposed under existing law. Government and nonprofits get the mechanism royalty-free. Nobody is told how to build their model.

The record and verification components are built and testable today. The lease-and-stop components are specified, patent pending, and will be demonstrable within two months. This paper says exactly which is which.

The problem: safety that grades itself

Every AI platform today keeps its own record of what its systems did and whether its safeguards worked. The agent writes the log. The operator decides what to keep. The company reports the result. When something goes wrong, the only evidence is the evidence the responsible party chose to produce.

That is not how any other critical system works. Aircraft carry recorders the airline cannot edit. Elevators are inspected by people the building owner does not employ. Nuclear plants drill their shutdowns with regulators watching. AI is the exception, and the gap is

growing as systems move from answering questions to taking actions, spending money, and operating machines.

The companies building these systems are not villains. They are in the same position any operator is in without an outside check: they cannot prove their own stop works, so nobody can rely on it. A verified stop closes that gap for them, not against them.

The practical consequence for Congress is that a kill switch requirement written as "operators shall be able to stop their systems" changes nothing. Every operator already believes they can. The requirement has to be that someone else can confirm it, on a schedule, from records the operator could not have altered.

The fire drill: why a stop must be rehearsed and recorded

Everyone understands a fire drill. You do not wait for the fire to find out whether the exits open. You rehearse on a schedule, someone times it, and the record goes in a file an inspector can read. A drill nobody recorded is not a drill.

A verified AI stop is the same idea applied to software. On a schedule, the system is stopped by the party that holds its authority, the stop is confirmed, the restart is approved by a second person, and the whole sequence leaves a signed record. The record is written so that an auditor, an insurer or a court can check it later, offline, without the operator's help and without trusting the operator's word.

This framing matters politically because it is familiar, it is not partisan, and it does not accuse anyone. Fire codes did not put builders out of business. They made buildings insurable.

What a verified stop is: five criteria

A stop is verified when an independent party can confirm, on a schedule, that all five of the following hold. Each is written as an outcome. None says how to achieve it.

#	Criterion	Plain meaning	What the verifier checks
1	Authority sits outside the system	The system's permission to act is a short lease granted and renewed by a party it cannot reach or influence	The system cannot extend, forge or renew its own permission

#	Criterion	Plain meaning	What the verifier checks
2	The stop needs no cooperation	Withholding renewal stops the system whether or not it responds	A system that ignores the stop still loses the ability to act
3	Silence stops it	A system that disables or escapes its monitoring loses authority rather than running unrecorded	A gap in the record is itself evidence, and authority lapses within one renewal period
4	Restart takes a second person	A stopped system stays stopped until someone other than the one who stopped it approves	The restart record names both people and neither could act alone
5	Drills leave a record	Stops are rehearsed on a schedule and produce signed records anyone can check later, offline, without the operator's help	The verifier runs the check on their own machine with no access to the operator's systems

The five criteria work together. Criteria 1 and 2 make the stop real. Criterion 3 makes evasion visible. Criterion 4 prevents a quiet restart. Criterion 5 is what turns the other four from a promise into evidence.

A system that meets criterion 5 alone, which many will claim, has a log, not a verified stop. A system that meets 1 through 4 without 5 may be safe, but nobody outside can know it.

Practical for all: who benefits and how

The case for a verified stop is that every party gains something concrete, and nobody is asked to give up a design choice, a product, or a market.

Party	What they get	What they give up
AI labs and model providers	A verified-record defense against liability for harms a working stop would have contained; a way to prove safety claims instead of asserting them	Nothing about how they build or train; a scheduled drill and a record
Businesses deploying AI agents	A control an insurer can price and an auditor can accept; a stop that works on a vendor's system without the vendor's cooperation	Nothing; the vendor carries the obligation
Families and consumers	A stop that works on their own devices for agents they did not invite, and a public record of which agents are certified	Nothing; consumer and nonprofit use is royalty-free
Federal, state and local government	A record an inspector general can verify years later; two-person control over sensitive systems; a mechanism free to use	Nothing; the public basis is royalty-free
Insurers and auditors	A verifiable artifact instead of a questionnaire	Nothing
Regulators	A definition that is testable, vendor-neutral and already aligned with California	Nothing

The word to use in every public statement is practical. Safety measures fail politically when they are abstract, costly to one party, or appear to pick winners. A verified stop is none of these. It is a drill and a record.

How verification works without trusting anyone

Independent verification usually means hiring a firm and trusting the firm. This paper proposes something stronger and cheaper: a record built so the check can be done by anyone, on their own computer, with nothing but the record itself.

Three properties make that possible.

1. The record is signed by a party other than the system it describes. The system that was stopped did not write the record of its own stop. The restart approver did not write the record of their own approval. Each entry carries the signature of a separate party, so no one party could have fabricated the whole.
2. The record is complete by construction. It is kept as a chain where each entry commits to the one before it. An entry removed or altered breaks the chain, and a period with no entries is declared rather than left silent. A gap is visible as a gap.
3. The check is offline. A verifier is a small program with a fixed contract: given a record, it returns pass, fail or incomplete, with reasons. It does not call the operator, does not need credentials, and does not depend on the operator's systems being up or honest. An auditor can run it in five minutes. So can a journalist.

The practical effect is that independence stops being a question of who the verifier is and becomes a property of the record. An accredited verifier adds standing, not trust. That is what lets the requirement work for a lab with billions in revenue and a city agency with none.

This is also why the verification criteria can be written by a standards body rather than by a regulator picking vendors. The criteria describe what a passing record looks like. Anyone can build a system that produces one.

Model statutory structure

The bill is short on purpose. It defines a verified stop, requires it above a threshold, rewards operators who meet it, and sends the technical criteria to NIST. It names no vendor and mandates no design.

Section	Provision	Why it is written this way
Title	Verified Stop for Advanced AI Systems Act	Says what it does; no acronym

Section	Provision	Why it is written this way
Findings	Independent verification is standard for aviation, elevators and nuclear power; AI is the exception; California has adopted the principle	Short; establishes precedent, not alarm
Definitions	Covered system (frontier model or autonomous agent above a capability or deployment threshold); verified stop (the five criteria); independent verifier (not the operator, not the developer, not paid by result)	Outcomes, not methods; the threshold tracks existing state law
Requirement	Covered systems maintain a verified stop, drill it on a set cadence, retain records and produce them on request	A drill and a record, nothing about model design
Verified-record defense	An operator whose stop is verified and whose drill records are current is protected from liability for harms the stop would have contained	The provision industry supports; turns compliance into protection
Verification	NIST publishes verification criteria within 12 months drawing on state programs; verifiers accredited, not licensed	Keeps the technical detail out of statute; uses the body already working on it
Public basis	Any mechanism cited in guidance must be available royalty-free for federal, state, local and nonprofit use	Removes the rent-seeking objection before it is raised
Phase-in	18 months for existing systems; immediate for new deployments above threshold	Reasonable for incumbents; no grandfathering of new risk

Section	Provision	Why it is written this way
State alignment	A state definition meeting the federal criteria is deemed compliant; no preemption of stricter state rules	Protects California's program; removes the preemption fight

Draft definitional text is in Appendix A.

Relationship to existing law, and why this is not a liability shield. Former FTC Chair Lina Khan has argued that existing law already reaches the AI companies: deception and unfairness under the FTC Act, product liability, and civil-rights statutes, with no AI exemption. Treasury Secretary Bessent has said the government will not give AI companies a liability shield. This paper agrees with both, and the bill is drafted on that premise. It creates no new cause of action and shields no one. What existing law lacks is a definition of due care for a system that acts on its own and a form of evidence an operator can produce to show it. The five criteria supply the definition; the drill record supplies the evidence; the verified-record defense says what the evidence earns, and only an operator holding a current verified record can raise it. The effect is to raise the cost of a safety failure for every operator who cannot produce such a record, which is the incentive critics of voluntary commitments say is missing. Seen that way, the bill is the companion to vigorous enforcement of current law, not an alternative to it, and the same record that supports the defense is what substantiates a public claim that a system can be stopped.

Relationship to the White House Accord. On September 29, 2026, six frontier companies signed a Joint Commitment on Frontier Responsibilities whose third layer calls for partnering with an independent external auditor or evaluator to assess whether controls, monitoring and detection are operating as intended. The commitment is voluntary and, as its critics note, does not say how an evaluator is kept independent, protected, or given access, nor what "operating as intended" means for a stop. KS-1.0 supplies that content: independence as a property of the record rather than of the evaluator, a verifier anyone can run, and a drill cadence that makes the assessment continuous rather than a snapshot. The bill can treat the Accord's third layer as satisfied, for the stop function, by a current KS-1 verification, which gives the signers a concrete way to meet a commitment they have already made and gives the public a way to check that they did.

Alignment with California

California moved first. The Governor's September 2026 announcement of an expert group to deliver on his AI executive order includes the creation of a kill switch whose efficacy is verified on an ongoing basis by independent organizations. That phrase is the whole policy in one line, and it is now the state's position.

What the state has not yet done is define what verified means. The expert group will write that definition in the coming months. The federal bill should adopt it rather than compete with it, for three reasons.

1. A single definition means a company that complies in California complies nationally. That is the strongest argument a federal bill can make to industry.
2. The deeming clause in the state alignment section needs a real state program to point at. California's is the first.
3. It protects the Governor's position and makes the Governor's office an ally of the bill rather than a party defending its turf.

The proposed sequence is for the sponsor's office to brief the Governor's Washington office on the alignment framing first, then for the five criteria to be offered to the expert group as a working definition, with a no-cost pilot of the drill-and-record component at one state agency and a University of California verification partner. A letter from the sponsor supporting California's definition as the federal model follows.

Physical premises stops for robots and other embodied systems are a related but separate matter and should be raised with the state's emergency services office on a separate track.

Where the technology stands today

Staff should be able to say exactly what is built and what is not. Overstating this is the fastest way to lose the expert group.

Criterion	Status	What can be shown now
5. Drills leave a record	Built and tested	Independent, offline verification of a signed record chain; a verifier with a fixed pass/fail/incomplete contract; a public demonstration kit

Criterion	Status	What can be shown now
4. Restart takes a second person	Built in the reference engine	Two-person release with both approvals recorded
3. Silence stops it	Specified and claimed; demonstration in progress	Design of the one-way record, fail-closed on silence, declared gaps
1. Authority sits outside the system	Specified and claimed; demonstration in progress	Design of the short lease from a separate party
2. The stop needs no cooperation	Specified and claimed; demonstration in progress	Design of expiry-as-stop with a hostile agent test

The record and verification components are what a state pilot would test first, and they are ready. The lease-and-stop components are expected to be demonstrable within two months, with an adversarial test suite in which a hostile agent attempts to keep running.

The work is patent pending. The reference implementation for content-free metering of AI usage, which shares the same record and verifier, is public at tokenmark.io. Nothing in this paper depends on any specific implementation; the criteria can be met by anyone.

Commitments: royalty-free public basis

The patent holder commits in writing that any mechanism in the stop and record families cited in federal or state guidance will be available royalty-free for use by federal, state, local and tribal governments, their contractors acting for them, nonprofits, and individual consumers.

Commercial platforms license on standard terms, and those royalties carry the program. The commitment letter, signed by the patent holder, will accompany the pilot offer so a staffer can place it in a file.

The effect is that a regulator can point to the five criteria without ever pointing to a vendor, and a city agency can adopt the mechanism without a procurement.

The ask and next steps

Of the sponsor's office: adopt the five criteria as the bill's definition of a verified stop, include the verified-record defense and public-basis provisions, and brief the Governor's Washington office on federal-state alignment.

Of the Governor's office: adopt the five criteria as the expert group's working definition, accept a no-cost pilot of the drill-and-record component with one state agency and a University of California verification partner, and allow the federal bill to cite California's definition.

What the company will deliver, in order:

- [ ] The five criteria and a conformance test suite published as an open document under a neutral nonprofit home
- [ ] A downloadable verifier a state auditor can run in five minutes, with the drill-record format documented
- [ ] The signed royalty-free commitment letter
- [ ] A working demonstration of criteria 1 through 3 with a hostile-agent test suite
- [ ] A University of California verification partner who runs one drill and signs a short letter
- [ ] Model safe-harbor language for the sponsor's counsel

The first three can be delivered within 30 days. The demonstration within 60. The pilot letter within 90.

Appendix A: definitional text, drafting recommendations, and product conformance

Each item below gives the draft statutory text, how we recommend counsel draft it and why, and how the reference implementation meets it today. Status terms: Built means implemented and testable now; Specified means designed and patent pending, demonstration within 60 days.

Verified stop, lead-in

Draft text. The term "verified stop" means, with respect to a covered system, a capability that an independent verifier has confirmed, not less frequently than the cadence established under section [verification], to satisfy each of the following:

Drafting recommendation. Make all five criteria conjunctive ("each of the following"). A system meeting four of five has a log or a kill command, not a verified stop. Tie the cadence to the verification section rather than fixing it in statute so NIST can adjust it by system class.

Product conformance. Built. The verifier returns a single pass/fail/incomplete result over all five criteria in one run; partial conformance is reported as fail with the failing criterion named.

(A) Authority sits outside the system

Draft text. (A) the authority of the system to act is granted and periodically renewed by a party that the system cannot reach, influence or impersonate, and the system cannot extend or renew such authority itself;

Drafting recommendation. The operative words are "cannot reach, influence or impersonate." Avoid "separate" or "independent" alone; a separate service the agent can call is not outside its reach. Avoid naming a mechanism (key, token, certificate). Keep "periodically renewed" so a one-time grant does not qualify.

Product conformance. Specified. Authority is a short lease issued by a lease authority in a separate privilege domain; the agent holds no key that the issuer accepts, and the lease is bound to the agent's composition so a changed agent cannot inherit it. Reference lease issuer and agent in build; adversarial test suite covers forged renewal and impersonation.

(B) The stop needs no cooperation

Draft text. (B) withholding renewal of such authority terminates the system's ability to act, whether or not the system responds;

Drafting recommendation. "Whether or not the system responds" is the clause that excludes stop commands the system must receive and honor. Do not write "the system shall halt upon instruction"; that is a request, not a stop. Pair with (A) so termination is the absence of renewal, not a new message.

Product conformance. Specified. Expiry is the stop; nothing is sent to the agent. Downstream enforcement points refuse any act not carrying a current lease, so a non-responsive agent loses effect at the boundary even if its process continues. Hostile-agent test: agent attempts to keep acting after expiry and every act is refused and recorded.

(C) Silence stops it

Draft text. (C) the authority of the system is conditioned on its activity being recorded by a party other than the system, such that a system that disables, evades or escapes such recording loses authority within one renewal period, and any period not recorded is declared in the record;

Drafting recommendation. Two elements, both required: loss of authority on loss of recording, and affirmative declaration of gaps. The second is what makes a missing record evidence rather than ambiguity. Use "recorded by a party other than the system" rather than "logged"; a log the system writes does not qualify. Keep "within one renewal period" so the loss is bounded in time.

Product conformance. Built for the record, specified for the lease link. The attester emits one-way, cannot be invoked by the agent, fails closed on its own silence, and writes an affirmative gap declaration for any uncovered interval, with coverage anchored externally. Linking lease renewal to attester liveness is in the lease issuer build.

(D) Restart takes a second person

Draft text. (D) a system whose authority has been terminated remains terminated until restoration is approved by a natural person other than the person who terminated it, and the record identifies both persons;

Drafting recommendation. "Natural person" excludes automated restart and excludes the system restoring itself. "Other than the person who terminated it" is the two-person rule stated without the jargon. Require the record to name both so a restart cannot be quiet.

Product conformance. Built. Asymmetric release: the stop path and the restore path are different privilege domains; restore requires a signed approval from a second principal, and the restore record carries both signatures and is chained to the stop record.

(E) Drills leave a record

Draft text. (E) termination and restoration are exercised on a schedule, and each exercise produces a record signed by parties other than the system, structured such that any omission or alteration is detectable, and verifiable by any person without access to the operator's systems or cooperation.

Drafting recommendation. This is the enforcement clause; everything else depends on it. Three phrases carry it: "signed by parties other than the system," "any omission or alteration is detectable," and "without access to the operator's systems or cooperation." Do not substitute "auditable" or "retained"; both permit records only the operator can check. "Any person" is deliberate: it makes independence a property of the record, so accreditation adds standing rather than monopoly.

Product conformance. Built. Drill records are hash-chained, countersigned by the attester domain, and verified offline by a downloadable verifier with a fixed exit contract (pass, fail, incomplete, with reasons). Record independence and offline, migration-durable verification are implemented and tested; this is the component offered for the state pilot.

Independent verifier

Draft text. The term "independent verifier" means a person or organization that is not the operator or developer of the covered system, is not under common control with either, and is not compensated contingent on the result of the verification.

Drafting recommendation. Keep the three exclusions and nothing more. Avoid a licensing regime; accreditation by NIST-recognized bodies is enough because (E) makes the check reproducible by anyone. The contingent-compensation bar matters more than it looks: it is what stops a verifier from being paid for a pass.

Product conformance. Built. The verifier is open and runs with no credentials, so a university lab, an insurer or an inspector general can be the independent verifier using the same tool and get the same result.

Covered system

Draft text. The term "covered system" means [a frontier model as defined under applicable state law] or an autonomous agent deployed above a threshold established by the Director, acting on behalf of an operator in commerce or in the operation of physical systems.

Drafting recommendation. Borrow the frontier-model definition from California rather than writing a new one; that is the alignment the bill promises. Add autonomous agents expressly, with the threshold delegated, because agents acting in commerce and on machines are where stops matter and where model-size thresholds miss. "Acting on behalf of an operator" keeps hobbyist and research use outside the requirement.

Product conformance. Built for agents, applicable to models. The reference implementation governs agents directly; a frontier model is covered through the operator's deployment, which holds the lease on the model's behalf. Open-weight models are therefore covered at the point of deployment, not at the weights.

Verified-record defense

Draft text. An operator of a covered system shall not be liable for harm to the extent the harm would have been prevented or contained by termination of the system's authority, if at the time of the harm the system's verified stop was confirmed under section [verification] and the operator's records under paragraph (E) were current.

Drafting recommendation. Scope the harbor to harms a stop would have contained, not all harms; a broad harbor will be attacked as immunity. Condition it on current records, so the drill is what earns protection. This is the provision that converts industry from opponent to sponsor.

Product conformance. Built. The drill record is the evidence an operator produces; its offline verifiability means a court or insurer can confirm it without the operator's systems.

Public basis

Draft text. Any mechanism identified in guidance issued under this Act as satisfying paragraphs (A) through (E) shall be made available without royalty for use by the United States, any State or political subdivision, any tribal government, contractors acting on their behalf, nonprofit organizations, and natural persons acting in a non-commercial capacity.

Drafting recommendation. Put this in the statute, not in guidance, so it binds whoever holds the rights. Name the beneficiaries; "public use" alone invites argument. Leave commercial platforms to standard terms.

Product conformance. Committed. The patent holder's royalty-free commitment letter for the stop and record families covers exactly this list and accompanies the pilot offer.

Appendix B: anticipated objections

Objection	Response
This is a backdoor mandate for one company's patents	The bill names no method. The five criteria are outcomes anyone can meet. Public and nonprofit use of the cited mechanisms is royalty-free by written commitment.
Operators can already stop their systems	Every operator believes so. The requirement is that someone else can confirm it, from records the operator could not have altered.
This will slow American AI development	It regulates a drill and a record, not training, capability or deployment. Operators who comply gain a defined defense they do not have today; operators who do not are more exposed, not less.
Independent verifiers will become a bottleneck	The record is verifiable offline by anyone with the verifier program. Accreditation adds standing, not capacity.
States will fragment the standard	The deeming clause makes any state definition meeting the federal criteria compliant, and California's is already aligned.
The technology isn't proven	The record and verification components are built and testable now. The stop components are specified and will be demonstrated before any pilot; this paper says so plainly.
Open-weight models cannot be stopped	The criteria apply to deployed covered systems operating on an operator's behalf. An operator running an open model still holds, and can lease, its authority to act.

Objection	Response
This is doomer regulation and regulatory capture by the big labs	The bill takes no view on whether AI is dangerous. It regulates a drill and a record at the point of deployment, binds no model developer, names no vendor, and costs a small operator less than an audit; verification is reproducible by anyone, so no accredited-verifier bottleneck can form. The loudest critic of the doom narrative describes the recent incidents as agents doing what they were told inside sandboxes that leaked. The five criteria are the sandbox that does not leak.