Submission to the National Institute of Standards and Technology

AI Watchdog Alliance℠ · A National Initiative · standards@aiwatchdogalliance.org · 970-236-8228

October 8, 2026

Center for AI Standards and Innovation, and the AI Risk Management Framework program
National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, Maryland 20899

Re: Submission of KS-1.0, Verified Stop Standard for Advanced AI Systems (Draft 0.9.3), as candidate verification criteria for independently verified stop mechanisms

The AI Watchdog Alliance submits KS-1.0 for NIST's consideration as a reference in guidance on verifiable control of advanced AI systems, and specifically as candidate criteria for what it means for a stop mechanism to be verified by an independent party on an ongoing basis.

Why now. The State of California has called for AI kill switches whose efficacy is verified on an ongoing basis by independent organizations, and pending federal legislation would direct NIST to publish verification criteria within twelve months of enactment. Neither has yet defined "verified." KS-1.0 offers a definition written as five testable outcomes, with a record format, a verifier contract and a conformance test outline, so that NIST can evaluate it against existing frameworks rather than start from a blank page.

The White House Accord. On September 29, 2026, six frontier developers signed the Joint Commitment on Frontier Responsibilities, which calls for an independent external auditor or evaluator to assess whether controls, monitoring and detection are operating as intended. The commitment is voluntary and does not say what "operating as intended" means for a stop, or how an evaluator is kept independent. 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. A current KS-1 verification is a concrete way for a signer to show the commitment was met, and a way for the public to check that it was.

What the standard requires. A deployment has a verified stop when an independent verifier confirms, on a schedule, that: (A) the system's authority to act is a short lease granted by a party it cannot reach or influence; (B) withholding renewal stops the system whether or not it responds; (C) a system that disables or escapes its monitoring loses authority within one renewal period, and any unrecorded interval is declared in the record; (D) a stopped system remains stopped until a different natural person approves restoration, and the record names both; and (E) stops are drilled on a schedule and produce signed, chained records verifiable offline by anyone, with no access to the operator. Clauses 6 through 8 state these as normative requirements with the test a verifier performs for each.

Physical machines (Criterion F). Draft 0.9.3 adds a sixth outcome for embodied systems. A robot, vehicle or other actuated machine meets the verified-stop definition when an independent party can bring it to a safe stop from outside the machine, without the cooperation of the machine, its software or its network; when that stop is rehearsed on a schedule and leaves a record anyone can check; and when the machine's stop-assurance level, including none, is disclosed to the owner before operation. Criterion F names no mechanism. A machine that cannot be stopped from outside is not barred, it is labeled.

Relationship to NIST frameworks. KS-1.0 is built on NIST work and maps to it in clause 11: AI RMF (AI 100-1) Govern, Manage 2.3 and 2.4, and Measure 2 and 4; SP 800-53 Revision 5 controls AC-3(2), AC-5, AC-6, AC-12, AU-5, AU-9, AU-10, AU-12, CA-2(1), CA-7, CP-4, IR-3, SC-7, SC-12 and SC-13; FIPS 180-4, 202 and 186-5 for primitives; SP 800-57 for key separation; SP 800-63B for approver authentication; and SP 800-207 for the enforcement-point model. For generative systems, the AI 600-1 risks addressed are dangerous recommendations acted on by agents, confabulated actions on an operator's behalf, and value-chain integration where a downstream operator cannot bound an upstream model's authority.

What makes the approach distinctive. Independence is a property of the record, not of the verifier. Because every entry is signed by a party other than the system it describes, chained so omissions are

detectable, and verifiable by a published tool with a fixed pass, fail or incomplete verdict, any accredited party reaches the same result from the same inputs. This lets a regulator rely on independent verification without creating a verifier bottleneck, and lets a laboratory, insurer or inspector general perform the check with no cooperation from the operator.

Openness and intellectual property. KS-1.0 names no product and requires no particular mechanism; conformance can be achieved by any implementer. The text is licensed under Creative Commons Attribution 4.0. The underlying mechanisms developed by the Alliance's founder's company are patent pending; the Alliance has received a patent disclosure from the holder, UnlockFi, Corp., together with a written commitment that such mechanisms are available royalty-free to federal, state, local and tribal governments, their contractors, nonprofit organizations and individuals. Commercial terms are published by the patent holder on reasonable and non-discriminatory terms. Working-group members affiliated with a patent holder do not vote on adoption of covered requirements.

Status and alignment. Draft 0.9.3 is open for public comment through November 15, 2026. The five criteria have been provided to the offices of the sponsor and co-sponsor of the pending federal AI Kill Switch Act and to the Office of Science and Technology Policy for the statutory-definition work under Executive Order 14434, so that one definition can serve federal and state use.

Specific requests. The Alliance asks that NIST (1) consider KS-1.0 clauses 6 through 8 and Criterion F as candidate verification criteria in any guidance on verified stop or kill-switch mechanisms; (2) consider a National Cybersecurity Center of Excellence demonstration project on verified stop records, for which the Alliance will provide the reference verifier, test vectors and a university verification partner; and (3) include the Alliance in any stakeholder process on this subject. NIST staff comments would be welcome in the comment period or at any time.

Enclosed: KS-1.0 Draft 0.9.3. The patent disclosure and public-basis commitment are published with the standard on the Alliance's standards page.

Respectfully submitted,

John M. Ossenmacher
Director, AI Watchdog Alliance℠
949-238-0120 · standards@aiwatchdogalliance.org