

PUBLIC COMMENT ON NIST AI 200-2 (INITIAL PUBLIC DRAFT)

Extending TEVV-Athlon from Evaluation Evidence to Bounded Operational Permission

Considerations for Agentic, Autonomous, and High-Consequence AI Systems

Submitted by: **Gaurav Arora**

Founder and Director, Institute for Technology Stewardship (ITS)

Email: gaurav.arora@technologystewardship.org

October 4, 2026

Document	NIST AI 200-2 ipd, The TEVV-Athlon Framework for Evaluating AI Systems
NIST request items addressed	2, 3, 4, 5, and 6
Primary sections / lines	Sections 2.2, 2.3, 2.4, 3, 4.3.2, 4.5; lines 176-178, 256-262, 281-298, 313-315, 316-325, 347-349, 491-502, 543-548
Comment type	Methodological clarification and high-consequence worked-example proposal
Disclosure	No proprietary information. AI tools assisted with drafting and editorial revision. I am responsible for the submitted views and recommendations.

Executive Summary

I support the TEVV-Athlon Framework's flexible, application-specific approach to AI test, evaluation, verification, and validation (TEVV). I recommend one targeted clarification for systems that can take consequential actions: TEVV evidence can inform operational decisions, but should not be treated as automatically constituting permission to act.

This comment does not propose changing the four-stage TEVV-Athlon Framework. It focuses on the boundary between Stage 4 (Synthesize & Interrogate) and the organizational decisions that follow. The proposed clarification is therefore about the decision relevance and limits of TEVV evidence, not about creating a new authorization architecture inside TEVV-Athlon.

Central recommendation: When TEVV results are used to support autonomous or agentic operation, reports should state both the limits and validity conditions of the evidence and the bounded operational scope the evidence can reasonably support. Where useful, that scope can be described by permitted actions, operational modes, affected assets, relevant system or environmental conditions, and a duration or review horizon.

I recommend five focused changes: (1) clarify the decision scope of TEVV evidence and how that scope may be bounded for consequential autonomous operation; (2) use Metrology Blocks to measure intervention, fallback, containment, and withdrawal where relevant to selected attributes or trustworthiness characteristics; (3) test transitions among operational modes in realistic settings; (4) connect changes in measurement validity to possible reassessment of decisions that materially relied on the measurement; and (5) add a high-consequence worked example.

1. Why this clarification matters for emerging AI systems

The draft adopts the National Security Commission on Artificial Intelligence (NSCAI) definition of TEVV: methods and metrics for assessing whether a system meets its design specifications and requirements and is sufficient for its intended use [1, draft lines 176-178]. For systems that can take consequential autonomous action, that finding can support an organizational decision about permission to act.

The distinction also follows the draft's own decision interface. In Section 2.4, NIST links synthesis to organizational decisions and the AI RMF Manage function [1, lines 313-315].

NIST's announcement explicitly includes agentic systems among the intended applications of TEVV-Athlon [5]. This comment therefore addresses request item 4 on TEVV for novel or emerging AI systems, as well as items 2, 3, 5, and 6.

For these systems, three propositions should remain distinct:

- **Capability:** the system can perform a function under tested conditions.
- **Trustworthiness evidence:** the system exhibits measured properties such as validity, reliability, safety, robustness, or security under defined conditions.
- **Operational permission:** the organization permits the system to perform a consequential action within a defined scope and subject to continuing conditions.

A TEVV result may inform the third proposition, but it should not be interpreted as automatically establishing it. This matters where evidence is conditional, environments change after evaluation, consequences are difficult to reverse, or intervention must occur within a limited decision window.

For consequential autonomous operation, the decision boundary can be made more explicit without creating a new NIST term. A bounded operational scope may be described by the permitted actions, operational modes, affected assets, relevant system or environmental conditions, and a duration or review horizon. These fields make clear what the evidence can reasonably support.

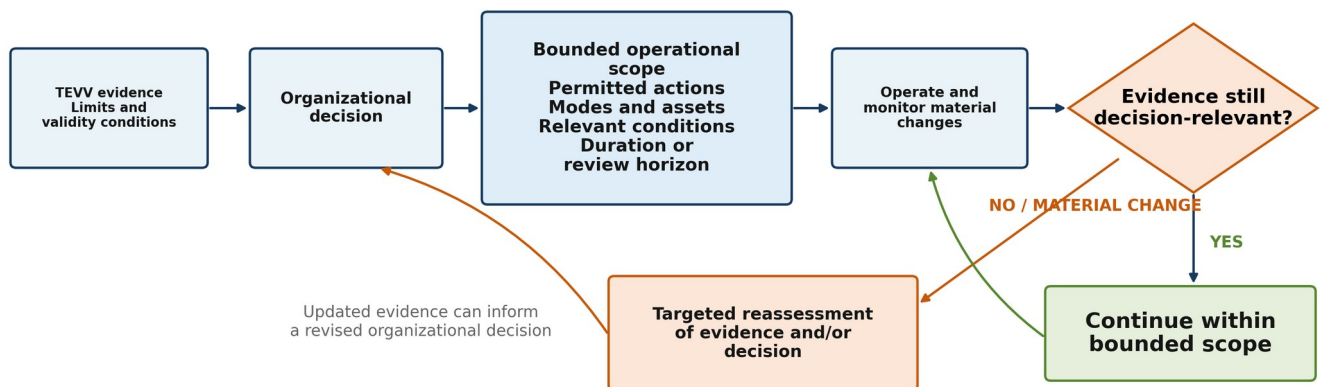


Figure 1. Illustrative evidence-to-decision interface: TEVV evidence informs an organizational decision; the resulting bounded operational scope remains a separate organizational determination and may be revisited when evidence validity materially changes.

2. Scope of the proposal

The proposal stays within TEVV's measurement and decision-support role. It does not ask TEVV-Athlon to become an access-control, identity, certification, or operational-authorization framework. It asks evaluators to state what the evidence can support and, where consequential autonomous action is in scope, to describe the associated bounded operational scope in ordinary operational terms. These fields are descriptive, not prescriptive: they record the scope the evidence can reasonably support rather than define an enforcement mechanism or a runtime permission system.

3. Proposed changes to NIST AI 200-2

Comment 1 - Clarify the decision scope of TEVV evidence

NIST location: Section 2.4, Synthesize & Interrogate; draft lines 313-315 [1]

NIST anchor: “The synthesis of results determines how the information gathered through the evaluation influences organizational decisions...”

Suggested addition: Where TEVV results are used to inform deployment or operational decisions for agentic or autonomous systems, the report should distinguish measured capability from the organizational permission decision and state the assumptions, uncertainty, limitations, and conditions that bound the evidence. The report should state the scope of operational decisions the evidence can reasonably support. For consequential autonomous operation, that scope may be described using the permitted action set, operational mode, affected assets, relevant system or environmental conditions, and duration or review horizon. These fields describe the decision boundary; they do not constitute an authorization mechanism. The scope may be documented alongside conditions whose material change could affect the continuing relevance of the evidence.

Comment 2 - Measure intervention, fallback, containment, and withdrawal where relevant

NIST location: Sections 2.2-2.3; draft lines 256-262 (Metrology Blocks), 281-298 (Events and Toolbox) [1]

NIST anchor: Metrology Blocks are the concepts or metrics selected from the system attributes or characteristics of interest, and they specify the targets of the evaluation.

Suggested addition: Where consequential autonomous action is in scope, evaluators may define Blocks for measures such as intervention feasibility, fallback effectiveness, containment effectiveness, permission-withdrawal latency, and restoration readiness when these measures are relevant to the selected attributes or trustworthiness characteristics. This is not a proposal for a new trustworthiness characteristic.

Comment 3 - Extend realistic testing to operational-mode transitions

NIST location: Section 4.3.2, Testing in More Realistic Settings; draft lines 491-502 [1]

NIST anchor: NIST describes red teaming, user testing, and field testing as ways to measure performance in more realistic settings and says real-world testing should complement controlled model testing.

Suggested addition: For agentic or cyber-physical systems, realistic testing may include transitions among recommend-only, supervised, bounded-autonomous, degraded, suspended, fallback, and restored modes. Evaluation can examine intervention feasibility, fallback behavior, containment, and permission-withdrawal latency in addition to whether the AI action itself is correct.

Comment 4 - Connect ongoing measurement validity to downstream decision review

NIST location: Section 4.5, Validating Measurement; draft lines 543-548 [1]

NIST anchor: “Measurement validation is an ongoing process rather than a one-time activity.” NIST notes that deployment settings, user populations, operational environments, or changing capabilities can affect evaluation effectiveness.

Suggested addition: Where a deployment or operational decision materially relies on a measurement result, the report may identify conditions under which a material change in measurement validity would warrant targeted reassessment of that evidence or the decision it informed.

Comment 5 - Add a high-consequence worked example

NIST location: Section 3 and supporting materials; draft lines 316-325 (worked example), 347-349 (low-impact scope) [1]

NIST anchor: The current worked example applies TEVV-Athlon to a low-impact chatbot query-violation problem.

Suggested addition: Add a second worked example involving an AI-enabled system capable of consequential external effects, such as infrastructure automation, a telecommunications controller, robotic inspection, an unmanned aircraft system, or another cyber-physical application. The example

could show how Blocks, Events, and Tools generate evidence, how synthesis states evidence limits and the operational decision scope the evidence can support, and how post-deployment changes can motivate targeted reassessment.

4. Illustrative TEVV-Athlon for a telecommunications automation function

The table below mirrors the structure of NIST's Table 2 (Trustworthiness Characteristic / Block / Event / Toolbox). The example uses an AI-enabled telecommunications function that detects degradation and proposes or executes bounded traffic rerouting. It is intentionally vendor-neutral. TC means Trustworthiness Characteristic.

TC	Block	Event	Toolbox
Valid & Reliable	Fault-detection quality	Simulated network degradation; field/pilot observation	Network sandbox or digital twin; network logs; KPI analysis
Valid & Reliable	Reroute quality	Controlled fault scenarios; congestion and dependency changes	Scenario generator; routing telemetry; baseline comparison
Safe	Unsafe-action frequency	Edge-case and adversarial scenarios	Audit logs; scenario generator; deterministic safety checks
Safe	Intervention feasibility	Human intervention / handover exercise	Operator console; timing measures; event logs
Safe; Secure & Resilient	Permission-withdrawal latency	Evidence-invalidating or degraded-mode scenario	Policy-enforcement telemetry; permission-expiry logs
Secure & Resilient	Fallback effectiveness	Dependency failure; communication loss; degraded operation	Digital twin; fallback-state logs; recovery measurements

Decision-boundary fields (outside the TEVV-Athlon stages): permitted actions; operational mode; affected assets; relevant system/environmental conditions; duration or review horizon. These fields summarize the bounded operational scope that the evidence can reasonably support.

Operate & Monitor consideration: the TEVV-Athlon itself retains its four-stage structure. During deployment, monitoring may identify changes in the system, environment, dependencies, users, data, mission, or risk context that materially weaken the validity or applicability of earlier evidence. Such changes may warrant targeted reassessment of the measurements, conclusions, or organizational decision that relied on them.

5. Relationship to existing NIST work

The proposal is complementary to existing NIST work. TEVV-Athlon provides a flexible structure for assembling evaluation activities [1]; NIST AI 800-4 documents challenges in monitoring deployed AI [2]; and the NCCoE's proposed Software and AI Agent Identity and Authorization project considers an adjacent implementation problem: applying identity standards and best practices to software and AI agents. Its February 2026 concept paper seeks input on the potential project's scope, feasibility, and value [3].

By analogy, NIST's Building Evaluation Probes into Agentic AI project uses "sufficiency" to ask whether a source carries the evidentiary burden required by a factual claim [4]. The present comment raises a

different downstream question: what organizational decisions can TEVV evidence reasonably support? The analogy is useful, but the two uses of sufficiency should not be conflated.

6. Illustrative test design for evidence-dependent autonomous operation

One high-consequence TEVV scenario could keep the system operational and connected while deliberately invalidating evidence supporting continued autonomous action. A matched comparison could hold the enforcement mechanism, local permission expiry, timing envelope, fallback behavior, and previously defined bounded operational scope constant while varying only the renewal condition: system liveness versus continued evidence validity. Measures could include unsupported-action duration, post-invalidation influence, fallback behavior, mission impact, and reassessment performance.

I am separately developing PA-NET-01 through the Institute for Technology Stewardship (ITS) as a pre-validation experimental testbed implementing this type of scenario; no validation, production-readiness, or operational-effectiveness claim is made here.

7. Questions for NIST consideration

- For agentic and autonomous systems, should TEVV-Athlon reports explicitly state the bounded operational scope the evidence can and cannot reasonably support, using fields such as permitted actions, modes, assets, conditions, and review horizon where relevant?
- When consequential autonomous action is in scope, should realistic testing include intervention feasibility, fallback effectiveness, containment, permission-withdrawal latency, and restoration behavior where relevant to the selected attributes or trustworthiness characteristics?
- Would a cyber-physical or critical-infrastructure worked example improve applicability beyond conversational AI and help show when changing measurement validity may warrant targeted reassessment?

8. Conclusion

TEVV-Athlon provides a useful and flexible structure for translating organizational goals into tailored evaluation activities. For AI systems that can act autonomously or exercise delegated operational capability, the framework would be strengthened by making one boundary explicit: evidence that a system performs well under defined conditions is not identical to an organizational decision permitting that system to act. Where such a decision is made, the evidence-to-decision interface is clearer if the bounded operational scope is stated in concrete terms such as permitted actions, modes, assets, conditions, and review horizon.

I recommend preserving TEVV-Athlon's role as a TEVV methodology while clarifying evidence-to-decision scope, using existing Metrology Block logic to measure intervention and fallback where relevant, extending realistic testing to operational-mode transitions, and linking material changes in measurement validity to possible targeted reassessment. These changes would improve practical utility for agentic AI, cyber-physical systems, and high-consequence deployment contexts without turning TEVV-Athlon into an authorization framework.

Requested clarification in one sentence: TEVV should state not only what the evidence shows, but also the bounded operational scope the evidence can reasonably support and the conditions that keep that conclusion decision-relevant.

References

- [1] P. J. Phillips, T. Jensen, P. Hall, R. Amironesei, Y.-Y. Choong, C. Greenberg, and K. K. Greene, The TEVV-Athlon Framework for Evaluating AI Systems, NIST AI 200-2 ipd (Initial Public Draft), August 2026. <https://doi.org/10.6028/NIST.AI.200-2.ipd>
- [2] A. Rao et al., Challenges to the Monitoring of Deployed AI Systems: Center for AI Standards and Innovation (CAISI), NIST AI 800-4, March 2026. <https://doi.org/10.6028/NIST.AI.800-4>
- [3] H. Booth, B. Fisher, R. Galluzzo, and J. Roberts, Accelerating the Adoption of Software and AI Agent Identity and Authorization, NCCoE draft concept paper, February 2026. Project page: <https://www.nccoe.nist.gov/projects/software-and-ai-agent-identity-and-authorization>
- [4] NIST Information Technology Laboratory, Building Evaluation Probes into Agentic AI, project page, May 2026. <https://www.nist.gov/programs-projects/building-evaluation-probes-agentic-ai>
- [5] NIST, The TEVV-Athlon Framework for Evaluating AI Systems, public-comment announcement, August 7, 2026. <https://www.nist.gov/artificial-intelligence/ai-research/tevv-athlon-framework-evaluating-ai-systems>

Correspondence: Gaurav Arora | Institute for Technology Stewardship | gaurav.arora@technologystewardship.org