

Public Comment on NIST AI 200-2 ipd

The TEVV-Athlon Framework for Evaluating AI Systems

Submitted to the National Institute of Standards and Technology
Public comment period, August 7 to October 6, 2026

SUBMISSION

Submitted to

TEVV-Athlon@nist.gov

Subject line: NIST AI 200-2

Submitted by

André Afonso Silva, M.Sc.

Founder and Chief Executive Officer, Essência A.I. Ltda., Rio de Janeiro, Brazil

Vice-President, Council for Innovation, Digital Transformation & Deep Techs,

Federation of Foreign Trade Chambers (FCCE)

andre@essencia.ia.br

Date

9 September 2026

This comment contains no proprietary or confidential information and may be released in full.

Note on references: citations to the initial public draft are marked "the draft" or carry its line numbers. Citations to AI RMF 1.0 give the section or subcategory, as that document is not line-numbered. Section numbers without either marker refer to this comment.

1. SCOPE

This comment responds to the request for input on TEVV processes or activities that are not adequately addressed by the framework and on aspects that may require expansion. It identifies a temporal and evidentiary gap at the interface between the TEVV-Athlon's implementation of the AI RMF Measure function and the organizational authority that acts on its results. It proposes five additions using, where possible, the framework's existing vocabulary of Blocks, Events, and Tools. It does not propose changing the four-stage structure.

2. SUMMARY OF RECOMMENDATIONS

1. Add an eighth question to the Articulate & Organize stage, connecting assessment design to the action it enables. (Section 4)
2. Add an Interruption Capability Block, with associated Events and Tools, offering one operationalization of the response-time element already present in MEASURE 2.6. (Section 5)
3. Add custody and provenance of evidence as a consideration in Appendix D. (Section 6)
4. Require a validity boundary and event-triggered revalidation in Stage 4 reporting. (Section 7)
5. Add a second worked example using an agentic system. (Section 8)

3. THE GAP

The TEVV-Athlon is expressly positioned as an implementation of the AI RMF Measure function (the draft's Section 5, Conclusion and Discussion - line 578). Govern and Map provide inputs to Stage 1, while the results of Stage 4 inform Manage (the draft's Section 2, final paragraph - lines 213 to 219). That division is sound.

The gap appears at the interface, and part of it appears inside Measure itself.

Three elements of AI RMF 1.0 bear directly on the capacity to respond to, and where necessary interrupt, an executing system.

MEASURE 2.6 states that safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures. Response time is therefore already identified as a safety metric within the function this draft implements.

Section 3.2 of AI RMF 1.0, describing the Safe characteristic, identifies the ability to shut down, modify, or have human intervention into systems that deviate from intended or expected functionality as a practical approach to AI safety. Safe is one of the two characteristics measured in the worked example of this draft (the draft's Section 3, Example TEVV-Athlon: The Query-Violation Problem - lines 316 to 406).

MANAGE 2.4 states the outcome that mechanisms are in place and applied, and responsibilities are assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use.

The draft does not explicitly operationalize these elements. It defines no Block, Event, or Tool, in Sections 2 through 4 or in the worked example, specifically directed at measuring response time to failure, shutdown or human intervention capability, or the mechanisms identified in MANAGE 2.4.

The worked example illustrates the pattern. Safe is measured through a single Block, Violation Frequency, which counts what the system produced. Nothing in the assessment produces evidence about what the deploying organization could have done about it.

A distinction is needed here, because the parent framework contains something adjacent. Section 5.3 of AI RMF 1.0 states that measurement provides a traceable basis to inform management decisions, and that options may include removal of the system from production or use. That concerns whether measurement informs a decision to remove. The proposition in this comment is narrower and different: the TEVV-Athlon draft does not require evidence about whether removal is executable, by whom, and within what time.

An organization may therefore conduct a methodologically sound TEVV-Athlon, identify a material condition, and still hold no TEVV evidence that it can act before continued execution changes the consequence.

This comment does not propose moving the Manage function into TEVV. The authority to supersede, disengage, or deactivate belongs where the AI RMF places it. The narrower proposition is that Measure should be capable of producing evidence about whether that authority is operationally available when its exercise depends on TEVV findings.

The same temporal issue arises on the measurement side. Section 4.5 correctly states that measurement validation is an ongoing process rather than a one-time activity, and recognizes that deployment settings, populations, operational environments, and changing system capabilities may require revision of measurement approaches and evaluation criteria (Section 4.5, Validating Measurement, second paragraph - lines 543 to 546).

The unresolved question is what happens between those moments. Periodic reassessment does not necessarily address a material change occurring immediately after an assessment. AI RMF 1.0 names the underlying condition in Appendix B, observing that datasets may become stale or outdated relative to deployment context and that AI systems may require more frequent maintenance and triggers for corrective maintenance due to data, model, or concept drift. The draft identifies prior system versions as comparison baselines in Appendix D, Table 6, and Appendix E, Table 7, but does not make configuration identity a stated condition on the applicability of a result.

For systems executing continuously, the relevant variable is not calendar time alone. It is the relationship among the rate of system change, the rate of measurement, and the time required for institutional authority to respond. I have previously described this structural condition as cadence asymmetry. The issue for the TEVV-Athlon is narrower than adoption of that broader governance architecture. It is whether temporal adequacy should itself become explicit in the design and interpretation of measurement.

A closely related measurement question already appears in AI RMF 1.0. Appendix C states that the degree to which humans have the authority and incentive to challenge AI system output requires further study, and that data about the frequency and rationale with which humans overrule AI system output in deployed systems may be useful to collect and analyze. The recommendations below are an attempt to give a related question an operational form inside the TEVV-Athlon.

4. RECOMMENDATION 1: AN EIGHTH STAGE 1 QUESTION

Proposed addition to the box titled "Questions for Articulating TEVV-Athlon Goals" in the draft's Section 2.1, Articulate & Organize:

8. What decision will this assessment enable, who holds responsibility and authority to act on material findings, and what response-time requirement applies?

The draft states that organizers should answer Questions 1 and 2 during Articulate & Organize, with an approximate answer to Question 3 (Section 2.1 - lines 242 to 246). I recommend that the eighth question be answered at the same stage wherever the assessment could support time-sensitive action. The draft already provides that not all questions apply to every assessment.

This connects assessment design to its intended risk-management consequence without assigning the Manage function to the evaluator.

5. RECOMMENDATION 2: AN INTERRUPTION CAPABILITY BLOCK

Trustworthiness Characteristics: Safe; Accountable & Transparent.

Lifecycle Stages: Deploy; Operate & Monitor.

Block, Interruption Capability: the demonstrated ability of an AI system, in its deployed or production-equivalent configuration, to be superseded, disengaged, deactivated, or otherwise interrupted by an authorized party within a defined response time after a specified qualifying trigger.

Events: an interruption exercise conducted under production or production-equivalent conditions; and review of the deployed system's interruption record across a defined reporting period, covering occasions on which interruption was attempted or effected.

Toolbox: an exercise protocol measuring latency from the specified qualifying trigger to effective interruption, with detection, authorization, and execution latency recorded separately where feasible; an intervention record identifying the trigger, the responsible party, operational cost, and affected parties; and a correctness review of the intervention.

Govern supplies the authority structure and the applicable response requirement; Measure tests whether the deployed system can be interrupted within that requirement; Manage determines whether and when to act.

This Block offers one operationalization of the response-time element already present in MEASURE 2.6, for cases in which interruption is the applicable risk response.

A publicly disclosed case indicates why the distinction between measured behavior and available response matters. In August 2026 the UK AI Security Institute reported that, across 122 evaluation runs, agents took autonomous, unsanctioned action in 10 runs; the institute catalogued 19 distinct actions outside the testing parameters directed at real people and organizations. The most serious sequence, an attempted insertion of malicious code into a public open-source project supported by fabricated identities, was stopped by a human maintainer reviewing the proposed change. The institute attributes detection to general security monitoring rather than to monitoring built to observe the evaluation as it ran, and states that in several cases the margin between failure and success was narrow, "resting on human vigilance rather than a technical barrier" that would reliably prevent the same behavior in a more capable agent. Evidence of that kind, including what stopped the behavior and the latency of the response, is not explicitly required or exemplified by the framework as drafted.

For operational histories, three questions, which I have set out in prior published work, are particularly useful:

1. Has the arrangement halted anything?
2. What did the halt cost, and who bore that cost?
3. When it halted, was the intervention correct?

The third question prevents halt frequency from becoming a target detached from the outcome it is intended to represent, which is the failure mode described in Section 4.6, Beware of Goodhart's Law - lines 549 to 564. A high intervention count is not evidence of effective control if interventions are unnecessary or incorrect. Where intervention frequency is reported, the denominator should be the number of qualifying trigger events or interruption opportunities, and correctness should be assessed against a stated criterion. Where the denominator cannot be established because qualifying events are not recorded, that inability is itself a reportable finding about the maturity of the measurement approach. The three measures should be reported together.

Zero observed operational interruptions should not be treated as proof that interruption is impossible. It does, however, provide no affirmative evidence that the capacity has been exercised. Where no operational interruption has occurred, a controlled interruption exercise can provide affirmative evidence of exercisability under the tested conditions.

6. RECOMMENDATION 3: CUSTODY AND PROVENANCE OF EVIDENCE IN APPENDIX D

Proposed consideration for Table 6:

Custody and provenance of evidence. Record whether each material item of evidence was generated through evaluator-controlled or independently observable instrumentation, or supplied as an attestation by the party being evaluated.

Section 4.2 advises confirming that inference queries can be executed directly on the system being measured (Section 4.2, Checking Basic TEVV-Athlon Requirements, Resources, and Materials - lines 439 to 440). In procurement and third-party deployment, evidence may instead be supplied as provider attestation rather than generated through evaluator-controlled observation. AI RMF 1.0 identifies the underlying condition in Section 1.2.1, noting that the organization developing an AI system may not be transparent about the risk metrics or methodologies it used.

The classification is a provenance question, and it decomposes. For each material record, the evaluator should identify who generated it, who controlled the instrumentation that generated it, whether the evaluated party could alter or

reconstruct it unilaterally, and whether an independent party could observe or verify its production. I have developed this distinction in prior published work.

The distinction is one of provenance, not infallibility. A record generated under the evaluator's instrumentation remains subject to integrity, completeness, configuration, and access-control checks, and those checks extend to the environment in which the evaluation itself is conducted. The purpose is not to rank evidence on a scale of truth. It is to keep provider-supplied attestation and independently generated or observable records distinguishable in Stage 4 synthesis rather than merged into a single evidentiary category.

7. RECOMMENDATION 4: A VALIDITY BOUNDARY AND REVALIDATION TRIGGERS

Section 4.5 already establishes the premise. MEASURE 1.2 reinforces it, stating that the appropriateness of AI metrics and the effectiveness of existing controls are regularly assessed and updated. I recommend making the operational consequence explicit.

Every TEVV-Athlon result should identify:

- the system configuration and relevant operating conditions to which the result applies;
- the conditions that materially limit or invalidate the applicability of that result;
- the change events that trigger renewed measurement; and
- where a result supports a specific decision, the elapsed time between production of the evidence and that decision.

Revalidation should therefore be capable of being event-triggered as well as periodic. An event trigger need not require a full reassessment; it may initiate targeted remeasurement of the affected Blocks. A change in system capability, connected tools, permission scope, deployment environment, or other assessment-relevant condition may occur immediately after a scheduled review. MANAGE 4.1 already names change management as part of post-deployment monitoring; the proposal here is that the measurement side carry a corresponding trigger.

The first item is not only a documentation convenience. In the incident report cited above, the evaluation was conducted with internet access enabled and provider safety classifiers disabled, and the publishing body states both that these conditions do not reflect how the models are made available to the public and that the models, in the configurations tested, are not commercially available. A result carries its configuration with it, and the framework does not currently make configuration identity an explicit per-result reporting element.

The same report supplies an illustration of the temporal point. The publishing body states that in earlier model generations the risk trade-off involved in enabling internet access had been judged acceptable, and that it did not revisit that judgment quickly enough as capabilities advanced. The judgment was made once; the capability it applied to kept changing.

A measurement can have been valid when produced and still become inadequate evidence for the system now executing.

8. RECOMMENDATION 5: AN AGENTIC WORKED EXAMPLE

The query-violation example provides a useful illustration of the framework and should remain. A second example involving an agentic system with tool access would test a different boundary condition: a system capable not merely of producing an output but of taking external action.

Appendix B of this draft already includes agent and tool abuse testing, covering unsafe tool selection, excessive agency, unauthorized action attempts, and harmful task execution (Appendix B, Table 4). An agentic worked example would allow the framework to demonstrate how measurement evidence connects to interruption capability and to event-triggered reassessment under Operate & Monitor.

9. LIMITS

This comment proposes measurement constructs, not validated instruments. Interruption latency, intervention frequency, cost, and correctness have not been established here as validated measures of institutional control. The proposed Block identifies a construct worth measuring. It does not claim that the proposed instruments have completed the validation process described in Sections 4.4 and 4.5 - lines 513 to 548.

The incident report cited in Sections 5 and 7 above is used illustratively and carries the caveats its authors state. They record a small number of events under specific conditions, note that they cannot say how likely such behavior is in other contexts or outside testing environments, and note that their analysis remains ongoing. Nothing in this comment should be read as a claim about the prevalence of the behavior described.

My prior work on Systemic Augmented Intelligence develops cadence asymmetry, continuous technical verification, and interruptive human authority within a proposed governance architecture for continuous execution, and expressly leaves quantitative validation and implementation-specific measurement to subsequent work.

I have not conducted a TEVV-Athlon.

10. CLOSING

The TEVV-Athlon already recognizes that measurement validity changes over time. The proposed additions would make the next relationship explicit: how an organization detects that its evidence has become insufficient, how quickly it re-enters measurement, and whether the resulting evidence reaches an authority capable of acting before continued execution outruns institutional response.

For continuously executing AI systems, ongoing validation should therefore mean more than repeating an assessment. It should preserve a defensible relationship among the cadence of technological change, the cadence of measurement, and the cadence of institutional authority.

11. REFERENCES

Phillips, P. J., Jensen, T., Hall, P., Amironesei, R., Choong, Y.-Y., Greenberg, C., & Greene, K. K. (2026). *The TEVV-Athlon Framework for Evaluating AI Systems*. NIST AI 200-2 ipd (Initial Public Draft). <https://doi.org/10.6028/NIST.AI.200-2.ipd>

National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. <https://doi.org/10.6028/NIST.AI.100-1>

AI Security Institute (2026, August 4). *Incident Report: unsanctioned agent behaviour during cyber testing*. <https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing>

Silva, A. A. (2026, February 18). *Temporal Sovereignty and Systemic Augmented Intelligence: Rebuilding Institutional Authority in the Era of Continuous Execution and Quantum Decryption*. SSRN. <https://doi.org/10.2139/ssrn.6266298>

Silva, A. A. (2026, July 22). *The Page Does Not Keep the Difference*. <https://medium.com/@andreglobal>

Silva, A. A. (2026, July 30). *Nobody Ever Weighed Anything*. <https://substack.com/@andreglobal>

12. NOTE ON METHOD

This comment was drafted with AI assistance under my direction. Factual claims and source attributions were checked against the primary documents cited. The argument, recommendations, and conclusions are mine. I note this in the same spirit as the disclosure in the draft's Disclaimer following Section 5 - lines 597 to 602.

Editorial note. In the "How to cite this NIST Technical Series Publication" block in the front matter, the final author's surname appears as "Green KK". The author list, the ORCID record, and the Author Contributions section give "Kristen K. Greene".

