

October 6th, 2026
National Institute of Standards and Technology (NIST)
100 Bureau Drive
Gaithersburg, MD 20899

Re: Feedback on NIST's Public Draft on the *TEVV-Athlon Framework for Evaluating AI Systems*

The undersigned civil society organizations appreciate the opportunity to submit comments in response to the National Institute of Standards and Technology (NIST) Request for Input on the first draft of the *TEVV-Athlon Framework for Evaluating AI Systems*. We commend NIST's leadership efforts in maturing the science, validity, and reproducibility of artificial intelligence (AI) evaluation, and acknowledge the time and consideration from the contributors that realized the public draft. We hope our comments can serve to help strengthen and reinforce elements of this important work.

Overview

Our overall view of the draft is that NIST has crafted and formalized the structure of general evaluation processes across a broad range of system types well. We understand the TEVV framework to encompass, at an intentionally higher level of abstraction, the general processes and considerations of evaluation, and acknowledge the intention to keep guidance streamlined. The current draft achieves this, particularly when it comes to evaluation of statistical, machine learning, and large language models (LLM) with a variety of degrees of complexity and underlying mathematical techniques. We believe the Block-Event-Tool structure to be a useful conceptualization for organizations seeking guidance on best practices when it comes to conducting rigorous AI evaluations.

Our substantive feedback is therefore limited to the framework's portability for when the purpose of evaluation is not only to measure a discrete ex-ante concern, but to understand what the important qualities to measure are. We refer to this paradigm here as discovery-oriented evaluations, where the purpose of the evaluation is to surface failure modes for a system class.¹ Our comments respond primarily to Questions 1, 3, 4, 5, and 6 of the Request for Input, and the questions each section addresses are noted in its heading to make the feedback easier to navigate. Our substantive feedback encompasses three points:

- (1) the definition of TEVV the draft adopts and the conformance orientation it carries;

¹ Burden, J., Tešić, M., Pacchiardi, L., & Hernández-Orallo, J. (2025). *Paradigms of AI evaluation: Mapping goals, methodologies and culture* (arXiv:2502.15620). arXiv. <https://doi.org/10.48550/arXiv.2502.15620>

(2) the framework's coverage of all types of AI system evaluation, including discovery-oriented evaluations, and the inclusion of an additional example demonstrating refinement and iteration loops within the framework; and

(3) further guidance for construct specification, given its outsized role in the framework's Define & Construct phase and the dependence of all downstream measurement upon it.

We try to offer concrete suggestions that the authors may consider, in revisions or companion work, if suitable, as well as our thoughts on promising directions in response to NIST's inquiry.

Substantive Feedback

1. TEVV definition and its sources (Question 1)

The draft adopts the definition of TEVV provided in the final report of the National Security Commission on Artificial Intelligence, under which TEVV comprises the methods and metrics used “...to determine that a technology or system satisfactorily meets its design specifications and requirements and is sufficient for its intended use” (lines 176–178).² This is primarily a conformance definition, in that it presupposes the existence of specifications and requirements against which a system may be assessed. There is some tension between that definition and the risk-management-oriented framing the TEVV-Athlon invokes in the same section, which sees AI system impacts serving as the primary driver of risk management activity and therefore NIST's TEVV supplying the evaluation process for identifying and measuring those impacts across the system lifecycle.

Evaluation undertaken to establish what a system's impacts may be will frequently proceed without specifications to conform to, and so the range of activities this TEVV evaluation framework encapsulates may not be best described as a determination of whether established requirements have been met, but rather as an assessment of how a system performs against classes of risks or threat vectors. **The authors may want to consider either broadening the adopted definition or stating that the framework is intended to accommodate evaluation beyond the NSCAI's conformance context so that entities are not led to read the narrower definition as a limit on the framework's orientation towards risk driven evaluation, or a limit on the value that TEVV activities can and should provide.**

2. Treatment of known unknowns and unknown unknowns in organizing the purpose and methods for an evaluation

² Phillips, P. J., Jensen, T., Hall, P., Amironeisei, R., Choong, Y.-Y., Greenberg, C., & Greene, K. K. (2026). *The TEVV-Athlon framework for evaluating AI systems* (NIST AI 200-2 ipd). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.200-2.ipd>

The draft, at present, does not distinguish between evaluations intended to surface information about known or anticipated risks and evaluations which are intended to surface risks that have not yet been identified by an organization. Making this distinction explicit as a starting point, perhaps in the *Articulate & Organize* stage, would help entities determine which arrangement of Blocks, and methods suit their purpose. The way TEVV methods should be used, and which are fit to purpose, will depend heavily on the organizer's starting posture towards what the evaluation is intended to accomplish. Confirmatory evaluation of a **known** risk would proceed from a well-specified Block, while evaluation of **unknown** or emergent risk may require discovery-oriented approaches before a Block can be defined at all. **We recommend that the authors consider naming both starting points in the framework and indicating how the stages may be used differently for each within the framework.**

3. TEVV-Athlon framework's modeling of iteration, refinement, and discovery-oriented evaluations (Questions 2 and 3)

Following from the previous point, AI evaluations are positioned as a tool that allows entities and organizations to trust the behavior and processes that an AI system will come to take on in their organizations and often, on their behalf. Agentic AI evaluations, for example, are increasingly not only concerned with understanding how a system performs against well-specified tasks or goals but also to building a more exploratory understanding of behaviors or proclivities of the system. While we believe that many of the processes involved in agent and multi-agent evaluation are captured by the TEVV-Athlon evaluation framework's discrete stages, the framework's center of gravity is most responsive to the assumption that failure modes are pre-specified and known by an organization beforehand. **We would therefore recommend the authors consider modeling iteration, refinement and discovery loops explicitly, especially for exploratory, discovery-oriented evaluations (such as those that may be particularly useful for agentic AI system classes).**

When the purpose of evaluation shifts from a confirmatory process of a particular task or feature towards evaluation as an exploratory practice, the mapping to the TEVV-Athlon evaluation framework can be less clear. The current draft does acknowledge that refinement can occur over these stages, but this could be made more explicit given that entities may turn to NIST's guidance, in part, to understand how messy, nebulous goals to advance understanding of AI system behavior in real world contexts can be given structure. This can be particularly useful in domains or use cases where relevant failure modes are less well understood, and TEVV activities are especially important to building an understanding that enables effective risk management.

NIST may consider addressing these points of feedback by demonstrating the framework's adaptivity through the addition of another example that is more complex. **Adding an additional, more complex evaluation use case alongside the Query-Violation Problem example would showcase the utility of the TEVV-Athlon framework across a broad range of evaluation**

paradigms that reflect common deployment realities.³ This additional example may be well suited for an exploratory evaluation use case that has iteration loops between stages. This can make it clear to practitioners that the TEVV-Athlon evaluation approach is more porous and robust to many forms of evaluation⁴ than entities may otherwise assume. The visual aids that accompany the written guidance can also make movement between the stages more apparent (refer to Figure 4).⁵

The framework's method appendices, while substantive, are similarly oriented towards the evaluation of an anticipated concern relative to the current text. The user and field testing methods collected in Appendix C—impact assessments, field pilots, focus groups and interviews, structured human-subject experiments, and usability research among them—are generally directed at how people come to understand and interact with a system already in operation. This could be more explicitly linked in text for entities pursuing the exploration described above, where the goal is to characterize general properties of a system class before the relevant failure modes have been named.

We agree with the draft's recognition that to be effective, AI evaluation must draw on human-centered methods from quantitative and qualitative social science research alongside technical ones, and with its description of validation as an ongoing rather than a one-time activity. Approaches drawing on these traditions are particularly well suited to surfacing risks, behaviors or failure modes that can then have evaluation designed to test more specifically. The authors may wish to consider including analytical case-based designs,⁶ qualitative⁷ and log-based analysis of agentic system behavior,⁸ abductive approaches that treat the anomalous observation as the starting point for explanation,⁹ and goal-free designs that withhold the intended objectives from the evaluator so that unanticipated behaviors and effects may be detected without contaminating the theory of outcome.¹⁰ The use of the latter may be more constrained and useful for those trying to understand latent multi-agent behaviors. We would recommend the authors consider representing similar approaches alongside the methods included in the existing appendices, whether within Appendix C or in a parallel treatment, so that the iteration and

³ Burden et al. (2025).

⁴ Burden et al. (2025).

⁵ Phillips et al. (2026).

⁶ Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Sage.

⁷ Sandelowski, M. (1995). Qualitative analysis: What it is and how to begin. *Research in Nursing & Health*, 18(4), 371–375. <https://doi.org/10.1002/nur.4770180411>

⁸ Kirgis, P., Kapoor, S., Rabanser, S., Nadgir, N., Ududec, C., Dubois, M., Allaire, J. J., Stosz, C., Hobbhahn, M., Steinhart, J., & Narayanan, A. (2026). *Log analysis is necessary for credible evaluation of AI agents* (arXiv:2605.08545). arXiv. <https://doi.org/10.48550/arXiv.2605.08545>

⁹ Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory*, 30(3), 167–186. <https://doi.org/10.1177/0735275112457914>

¹⁰ Youker, B. W., Ingraham, A., & Bayer, N. (2014). An assessment of goal-free evaluation: Case studies of four goal-free evaluations. *Evaluation and Program Planning*, 46, 10–16. <https://doi.org/10.1016/j.evalprogplan.2014.05.002>

discovery loops described above are accompanied by methods capable of supporting rigorous exploration-driven evaluations.

4. Guidance for construct specification is limited relative to guidance for downstream measurement of the evaluation construct (Questions 5 and 6)

The TEVV-Athlon framework locates construct formation within its early stages. Guidance available to entities on procedure tends to differ considerably across the stages. For example, the *Apply & Measure* stage is supported by an appendix that names and describes specific methods, which allows an organization to see the range of available approaches and to select among them with some understanding of what each affords. The *Define & Construct* stage, though it receives brief mention in Appendix E, does not have the same level of elaboration or method support.

Construct specification in AI evaluation practice does not yet have a reliably mature set of practitioner norms guiding best practices in the same way the measurement track has developed standardization through benchmark construction and other such methods. Current work is ongoing throughout the evaluation ecosystem to apply more rigor for construct development, however the area still lacks centralized best practices. NIST could be well suited to provide more initial guidance in this framework.

Blocks are described as the key concepts or metrics that an evaluation aims to assess, making it the unit that all downstream activities depend on. Entities that arrive at the *Define & Construct* stage with a well-specified quality already in hand will find the framework serves them well from that point forward. An entity that does not already know how to move from a broad trustworthiness characteristic to a defensible Block will have more difficulty. Industry teams tasked with evaluation may have less experience with the processes involved with forming a concrete Block target. The draft currently states that one must develop a Block without identifying or guiding users of the framework towards procedures of how to defensibly and rigorously define one before proceeding with subsequent stages.

We raise this since the consequences of an underspecified stage are disproportionate to the rest of the draft since failure here cascades into the other stages the current draft identifies. Rigor applied to a poorly specified Block produces numerical evidence that is internally consistent, well documented, and plausible—yet misleading due to construct misspecification at the *Define & Construct* stage. Where evaluation practice for a given Block may be mature and a surrounding ecosystem of benchmarks, harnesses, and established metrics already exists, an abstract framework layer will naturally be filled in by prevailing practice and what previous instantiations of the concept's operationalization exist in the literature. Where there is a paucity, however, or organizations find themselves in sectors with specialized tasks, the assumptions that underlie common Blocks may not port over unchanged, which can risk the validity of the subsequent results in a dissimilar context.

We recognize that NIST has good reason to hold the framework at a level of abstraction that does not constrain the methods entities elect to use, and we support that posture to ensure practitioner uptake. Extending support to construct specification would not require NIST to recommend a methodology, just that entities be shown that such methodologies exist for construct specification and elicitation and where they may be found as done for other measurement and validation stages in the current version. **We encourage the authors to consider developing a *Define & Construct* toolbox at parity with existing methods appendices (Appendices A-C), populated with established approaches to construct specification and elicitation.**

Several mature literatures may be well suited to this purpose that the authors could consider drawing from. Work in the social sciences on concept structure may provide guidance to practitioners on what steps they can use to reason through their own unique cases defensibly when using frameworks like the TEVV-Athlon which describes a process that moves from a basic concept, through secondary dimensions, to indicators. For example, construct mapping offers a documented procedure for moving from a qualitative description of a quality to an ordered set of observable indicators.¹¹ Where a Block would benefit from being constructed with input from stakeholders rather than derived from an existing literature, there is also work on structured elicitation methods including Delphi processes,¹² and through structured expert judgment that has already been adopted in scalable LLM-as-a-Judge approaches.^{13, 14}

We recognize that a full treatment of construct specification may exceed what the framework document can reasonably be expected to carry. A potential approach could be the model used for the AI Risk Management Framework, which retained a high level of abstraction while a companion resource carried more actionable detail.¹⁵ If the authors feel further support is necessary for the evaluator community, the authors might consider that model here, for dedicated supplementary treatment with a minimal set of named approaches retained within the framework itself as an appendix so that entities are not left without guidance for the *Define & Construct* stage in the interim.

Minor Feedback

5. Ecological validity across test and real world environments

¹¹ Wilson, M. (2023). *Constructing measures: An item response modeling approach* (2nd ed.). Routledge.

¹² RAND Corporation. (n.d.). *Delphi method*. Retrieved September 23, 2026, from

<https://www.rand.org/topics/delphi-method.html>

¹³ Goldfarb, R., Hall, A. B., Fisher, J., Salam, K., & Wilde, M. (2026). *Distilling expert judgment at scale* [White paper]. Forum AI. <https://byforum.com/whitepapers/distilling-expert-judgment-at-scale.pdf>

¹⁴ Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

¹⁵ National Institute of Standards and Technology. (n.d.). *NIST AI RMF playbook*. Retrieved September 23, 2026, from <https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook>

We appreciate the draft's brief mention of the gap between controlled testing and deployment. We believe that the draft could benefit from further elaboration on this point to support practitioners in how to judge whether results obtained in a test environment are likely to transfer to deployment and whether the assumptions made support external and ecological validity. While measurement validity is briefly discussed in Appendix D, mention of the affordances given to the testing environment itself as a relevant factor could be more prominently positioned in sections 4.2., 4.4., or 4.5, particularly given its relative importance to the evaluation of agentic system capabilities. Since trying to achieve similarity of the testing environment to generalize results to real world environments may also be cost-intensive, we recommend that the authors include this as a resourcing question in the initial deliberation guide offered in the *Articulate & Organize* stage.¹⁶

6. Mapping included methods directly to TEVV-Athlon framework stages (Questions 2 and 3)

At present most methods mentioned through the existing appendices are organized against stages of the AI system lifecycle rather than against the four stages of a TEVV-Athlon evaluation itself. A column mapping in the existing appendix tables showing where each method is best suited within the framework's own stages would help practitioners compose a TEVV-Athlon and learn to select and employ each method at the point in an evaluation where it is most useful. For instance, this could specify which methods would support construct specification in *Define & Construct*, tools for evidence collection in *Apply & Measure*, or approaches that are best for soliciting stakeholder interpretation in *Synthesize & Interrogate*. Depending on the methods, it may also be helpful to specify which actors should intervene at which stage, given that the TEVV framework relies on interdisciplinary teams with specialized domain expertise, and which actors should work in collaboration with each other in an iterative cycle of evaluation.

7. Specify the framework's scope and coverage of multi-agent systems (Questions 2 and 4)

The framework states the intention to cover a broad swath of AI system types. The framework could benefit from specific considerations for different types of AI systems such as models, prompt agents (i.e., a chatbot where there is a base model with decision logic, filters and classifiers, and data sources for retrieval-augmented generation), and agents with access to tools (i.e., a chatbot that has access to both internal and external data sources and can take actions). Different types of AI systems can present varying evaluation requirements and considerations at every level; considerations specific to the *Define & Construct* stage are particularly important due to their downstream effects on subsequent steps.

¹⁶ Phillips et al. (2026). p.5

For agentic systems, it is less clear at present whether multi-agent systems, in which several AI agents interact with one another, as well as with tools and their environment, are intended to fall within the framework's scope. It has been widely noted that evaluations intended to capture system-level interactions can naturally exhibit failure modes that arise from the interaction itself and this may not be initially observable when agents are evaluated on discrete tasks on an individual basis. To match the framework's ambition of coverage over the stated system types, representation of evaluation methods for agentic systems, such as analysis of tool calls, traces, could also be incorporated.¹⁷

If multi-agent systems are intended to be covered by the agentic system scope then guidance on how Blocks, Events, and Tools apply to interactions and evaluations between agents, or of the orchestration system as a whole, would be valuable for practitioners. If they are not then a clear statement of the framework's boundary, accompanied by clear classification with regards to complex system arrangements of AI systems that may be stated to be covered (e.g., LLM and agentic systems) would help entities understand where supplementary approaches are needed. In either case, clarity on scope would support the clarity of the framework's intended usefulness for novel and emerging AI systems.

8. Additional useful figures for the draft

The framework's terminology introduces several nested concepts e.g., trustworthiness characteristics, Metrology Blocks, Events, and the Tools that populate the Toolbox. Consider adding a hierarchical diagram showing how these terms nest to introduce the concepts, which would complement the written descriptions and assist entities encountering the vocabulary for the first time and improve the clarity. For example, how several Blocks instantiate different forms of evidence would be a useful relationship to demonstrate. Current figures could also benefit from more clarity for the viewer to follow such as Figure 4 in which the Metrology Blocks include *Reliability* and *Validity*, but the Table 2 above it show *Helpfulness* (under TC: Valid) and *Violation Frequency* (under TC: Safe). Having a direct mapping would help readers follow the figure from the table and understand where within the TEVV framework to apply the components.

Additionally, providing figures that help map the TEVV framework to an example AI development lifecycle figure could help readers understand where they might apply this framework in the stages of their own development.

9. Multi-cultural/multi-lingual implications

The framework currently assumes as default that the system is in an English-speaking domain and in a mostly Western/Global North context. It may be helpful to acknowledge the

¹⁷ Kehkashan, T., Abdullah, M., Al-Shamayleh, A.S. et al. From benchmarks to deployment: a comprehensive review of agentic AI evaluation. *Artif Intell Rev* 59, 167 (2026). <https://doi.org/10.1007/s10462-026-11571-0>

considerations that may need to be taken into consideration when evaluating in multi-cultural and multi-lingual contexts, in light of mass adoption of AI systems across the non-Western contexts. NIST has a unique opportunity to create guidance on engaging with language and subject matter experts to ensure their safeguards are robust across language and cultural norms to ensure that key safety risks are addressed for users outside of English speaking countries.

Promising Directions

10. Opportunity for participatory methods throughout (Question 5)

The draft's treatment of stakeholder involvement, particularly in Appendix C, often isolates participatory methods to the point where issues are identified. We would encourage the authors to consider whether that same involvement is warranted at later stages as well, so that participation is represented across both the framing and the validation of an evaluation.¹⁸ Judgments about whether a measured result is acceptable, where a threshold ought to be set, and what a given level of residual risk means for the people subject to a system are not purely technical determinations, and the parties best positioned to speak to them are frequently those the evaluation is intended to protect, particularly where a system is evaluated for real-world use. Extending participatory methods through the *Synthesize & Interrogate* stage, into the interpretation of findings and the setting of acceptance criteria, rather than into problem formulation alone, would make the framework's existing commitment to stakeholder input more consequential, and would be consistent with evolving evaluation traditions in which the standards of judgment are themselves established together with those affected. Employing such approaches in process similarly provides organizations with insight into the reception of their offerings by those impacted, and allows development to proceed with informed and well-validated product specifications.¹⁹

11. Domain-specific toolbox starter packs and “playbooks” leveraging the AI Metrology Center (Question 6)

The AI Metrology Center, in assembling metrics, methodologies and tools organized against the TEVV stages and measurement categories, is a very valuable complement to the framework and we appreciate NIST making it available as a resource. Entities working through a TEVV-Athlon will find in it a practical inventory to draw upon as they move from a defined Block toward the Events and Tools that might produce evidence for it. We see an opportunity for the two resources to be used together more deliberately since the framework's appendices direct practitioners here.

¹⁸ Schwartz, R., Westling, C., Briggs, M., Fadaee, M., Nejadgholi, I., Holmes, M., Rashid, F., Carlyle, M., Taik, A., Wilson, K., Douglas, P., Skeadas, T., Waters, G., Chowdhury, R., & Lacerda, T. (2026). *CIRCLE: A framework for evaluating AI from a real-world lens* (arXiv:2602.24055). arXiv. <https://doi.org/10.48550/arXiv.2602.24055>

¹⁹ Matias, J. N., & Price, M. (2025). How public involvement can improve the science of AI. *Proceedings of the National Academy of Sciences*, 122(48), Article e2421111122. <https://doi.org/10.1073/pnas.2421111122>

The authors may wish to consider developing domain starter packs, or sector profiles, in the manner of those developed for the AI Risk Management Framework,²⁰ that build upon the general guidance of this draft by carrying a small number of illustrative Blocks for a given domain, the Events and Tools suited to them, and the corresponding entries in the Center. Housing and lending, healthcare, and employment sectors, for example, would each be well-suited to support such a companion.

We are encouraged by NIST's decision to develop the TEVV-Athlon series for evaluation of AI systems and by the evident care taken by the authors in this first public draft. We would welcome opportunities to continue this conversation as the series develops, and we believe that listening sessions convening practitioners, civil society organizations, and evaluation researchers would be well received by the community, if useful. We thank the authors for their work and for the opportunity to respond.

Sincerely,

1. National Fair Housing Alliance
2. Center for Democracy and Technology
3. Center for Organizing
4. Interfaith Center on Corporate Responsibility
5. Sikh American Legal Defense and Education Fund (SALDEF)
6. Center for Media and Digital Governance at the Open Markets Institute
7. National Urban League
8. HTTP (Hispanic Tech & Telecommunications Partnerships)

²⁰ National Institute of Standards and Technology. (n.d.). *Use cases*. Trustworthy and Responsible AI Resource Center. Retrieved September 23, 2026, from <https://airc.nist.gov/airmf-resources/usecases/>