TECHNET
THE VOICE OF THE INNOVATION ECONOMY

1420 New York Avenue NW, Suite 825
Washington, D.C. 20005
www.technet.org | @TechNetUpdate

September 12, 2025

National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899

Re: Proposed Zero Draft for a Standard on AI Testing, Evaluation, Verification, and Validation

To Whom It May Concern:

TechNet appreciates the opportunity to comment on NIST's proposed zero draft for a high-level standard on AI testing, evaluation, verification, and validation (TEVV). Many of our nation's leading AI developers, deployers, researchers, and users are TechNet members and are utilizing industry-approved and science-backed TEVV methods and governance frameworks. To this end, TechNet appreciates NIST's constructive approach in creating an overarching framework to support AI practitioners in designing appropriate TEVV techniques for specific systems and cases, rather than mandating prescriptive TEVV methods.

TechNet is the national, bipartisan network of technology CEOs and senior executives that promotes the growth of the innovation economy by advocating a targeted policy agenda at the federal and 50-state level. TechNet's diverse membership includes dynamic American businesses ranging from startups to the most iconic companies on the planet and represents five million employees and countless customers in the fields of information technology, artificial intelligence, e-commerce, the sharing and gig economies, advanced energy, transportation, cybersecurity, venture capital, and finance.

TechNet supports NIST's effort to create a high-level, internationally relevant standard on AI TEVV. The rapid advancement and integration of artificial intelligence into critical sectors have created an urgent need for a standardized approach to TEVV. To help ensure the standard is both practical and effective, such a framework should provide a common language and methodology for developers, deployers, and users and offer actionable guidance that practitioners can implement with the flexibility to adapt to the evolving capabilities and emerging risks of next-generation AI systems. The following recommendations on the "Proposed Zero Draft for a Standard on AI TEVV" are intended to strengthen its foundation and help achieve its essential goals while ensuring it can serve as a durable and effective standard for managing AI risk.

Elevate Adversarial Evaluation

The draft acknowledges inconsistencies in the use of the term "red teaming" and aims for clarity. However, it currently relegates most specific TEVV methods, including adversarial approaches, to an appendix ("Appendix 2: Technical and sociotechnical approaches and methods"), treating them as part of an "incomplete catalog" rather than a core practice. Adversarial evaluation, including persona-based red teaming, should be elevated from the appendix to a core component of the TEVV process outlined in the main body. This section should define the practice, establish its importance for discovering novel risks beyond standard testing, and provide guidance on when and how it should be implemented as a core, rather than optional, step for high-risk systems. In particular, adversarial evaluation which involves deliberately trying to elicit problematic outputs from AI systems by feeding data most likely to cause the AI system to fail or produce material that may be unsafe, harmful, or offensive is foundational for evaluating the robustness of an AI system against problematic prompts or unexpected inputs. As an example, one could apply reinforcement learning to the selection of environmental factors as input to the test cases in a modeling and simulation-driven test bed that would include the AI software running the autonomous systems. The environment could be treated as a “thinking adversary” in an asymmetric game. The environment would learn optimal strategies to defeat the autonomous system, even as the autonomous system learns improved strategies to counter the environment. This would both maximize the chance of finding key vulnerabilities and weaknesses and help discover ways to mitigate or avoid those vulnerabilities.

In addition to making red teaming a core practice, there should be additional guidance on the collaborative function between offensive (red) and defensive (blue) security teams. TechNet recommends including purple teaming as an important component of adversarial evaluation, threat identification, and risk mitigation. Purple teams function differently from, and can be more complex than, red teams and blue teams due to their combination of both red and blue team functions. This section should define the concept as a collaborative exercise where red team findings are used to test and improve blue team detection and mitigation capabilities in real-time. Guidance should emphasize how purple teaming closes the loop, ensuring that identified vulnerabilities are not just documented but effectively addressed and that defenses work as expected.

Include Sector-Specific Examples and References

To make the framework more actionable, the appendices and main body should directly reference and provide examples using established, sector-specific resources and widely used AI security frameworks. Incorporating guidance on how to use frameworks, such as the MITRE ATLAS for adversarial threat modeling, OWASP Machine Learning Top 10 for common vulnerabilities, or SR 11-7 for model risk management in financial services, would provide practitioners with concrete tools to translate the standard into practice. Practitioners are also more likely to adopt a

new standard if it aligns with and builds upon the tools and methodologies they already use and that are more relevant for their sector-specific needs. By referencing and integrating the concepts from previously established, sector-specific frameworks, the TEVV standard can provide practitioners with a clear bridge from their existing compliance and safety obligations to the specific demands of testing and evaluating AI systems.

Expand Agentic AI Scenarios

A dedicated section should be added to discuss agentic AI. With rapid industry development, agentic AI introduces unique challenges and risks to the AI ecosystem given the autonomy provided to the AI system and its ability to take human-independent actions to achieve an objective. A separate approach for evaluating systemic risks resulting from agentic AI models being empowered to act on their own should involve guidance on evaluating long-term planning, goal alignment, and the potential for unintended consequences in open-ended environments. The framework should look to address how to test for emergent behaviors and ensure that such systems have reliable safety protocols. Validation measures could include testing autonomous agent's ability to safely interact with APIs, databases, external systems, and other AI agents as well as validating that the agent maintains intended objectives under adversarial pressure.

Enhance AI Test Automation

Generative AI and other machine learning techniques are now being used to improve and automate many parts of the testing process. Generative AI can automatically create comprehensive test cases and realistic, synthetic test data based on requirements and documentation. AI can be used to automatically adapt test scripts when user interface elements or code change to help solve the challenge of maintaining tests for rapidly evolving applications. It can also analyze historical data to predict where defects are likely to occur, allowing quality assurance teams to focus testing efforts on high-risk areas. Intelligent test orchestration can improve testing efficacy as AI can learn from past behavior and optimize test runs across various environments for improved speed and consistency. These are all examples of how test automation could be more thoroughly incorporated into the TEVV framework to help address the unique challenges of AI, which differ from traditional software due to its complexity and non-deterministic behavior.

Establish a Framework for Continuous, Lifecycle Spanning Monitoring

The draft mentions applying TEVV across the "AI systems lifecycle" and discusses "in-situ evaluations," but it does not explicitly recommend or encourage continuous monitoring *after* a system has been deployed. The framework should be updated to encourage continuous, runtime TEVV that goes beyond pre-deployment testing to include ongoing monitoring of the system's performance, behavior, and security in

its operational environment and further testing when key metrics change. Overall, testing should not be a final step, but rather a continuous process that is integrated into CI/CD (Continuous Integration/Continuous Deployment) pipelines to provide immediate feedback. This involves defining system objectives, translating them into measurable processes, and ensuring those processes are consistent and repeatable across multiple assessments. Dynamic evaluations should be considered alongside static testing to ensure that model drift, new vulnerabilities, and unexpected interactions are detected and addressed promptly. This will ensure the TEVV framework represents a dynamic, ongoing assurance mechanism that certifies a model remains safe, secure, and effective long after its initial deployment.

Strengthen Supply-Chain Verification Guidance

An AI model is not a monolithic entity; it is the product of a complex, multi-stage supply chain. A vulnerability or compromise in any single link of that chain can undermine the security, safety, and integrity of the final system. The draft correctly identifies the need to manage third parties and supply chains as a key challenge, and it notes that an organization's position in the supply chain is a key variable. However, this section could be strengthened by providing more specific recommendations on supply-chain verification. A sophisticated model can have its integrity undermined by a poisoned training dataset, a backdoored open-source dependency, or a compromised cloud environment. Robust supply chain guidance would ensure that security and safety considerations are built-in from the very beginning, preventing an organization from investing significant resources in testing a model that was already critically flawed due to a vulnerability inherited from its components. The guidance should detail the need for documentation covering a model's origin, training data, architecture, and any upstream components. This creates a "chain of custody" for AI models, allowing organizations to identify and mitigate risks inherited from third-party sources before they are integrated. It could also include vulnerability scanning, data quality audits, and vendor risk management strategies. By integrating supply chain verification measures, the TEVV framework would ensure that trust is not just assumed but is actively built and validated at every stage of the AI supply chain.

Refine the Concept Map

Refine the concept map to create a clearer distinction between evaluator intent and the system’s threat model. For example, the map could differentiate between TEVV activities aimed at *assurance* (verifying a system meets its stated requirements) and those aimed at *adversarial discovery* (finding unknown flaws). This would clarify that the choice of TEVV method should be driven by both the system's specific threat landscape and the goals of the evaluation. One way of refining the concept map could include a hierarchal diagram that is specific to the AI system and its context that defines the threat model for each AI system and then identifies evaluator intent and evaluation methods for each threat. This would ensure evaluators are explicit about whether they are confirming existing specifications or

searching for new, unknown flaws and make it easier to see gaps in the TEVV plan to allow for more comprehensive coverage. It would also create a shared vocabulary for developers, security testers, and project managers to discuss risk and evaluation strategies and lead to more actionable guidance in selecting the appropriate methods to test for risks based on clear strategic goals.

Develop a Standardized AI Metrics and Measurement Library

Organizations often default to simple metrics like accuracy, which fail to capture the full picture of a model's performance regarding fairness, robustness, or privacy. The TEVV framework should include a standardized library of recommended metrics for different AI risks and use cases. For TEVV to be effective, it should be grounded in **concrete, consistent, and context-appropriate measurement**. Simply stating that a model should be "robust" or "fair" is insufficient. There should be a common language and a standard set of tools to quantify *how* robust or *how* fair it is. A standardized metrics library would make TEVV results more consistent, comparable, and meaningful across different organizations and industries, moving beyond simplistic measures toward a more holistic and responsible evaluation of AI systems. It would also serve as a common language and a starting point for organizations by providing guidance on which metrics are most appropriate for specific applications. It would not need to be a rigid, one-size-fits-all checklist, but instead a comprehensive list of metrics from which organizations can select the most appropriate measures based on their specific context, risk tolerance, and legal requirements. In this way, the library would be structured around key risk categories, providing specific, mathematically defined metrics for each.
Such a library would provide a shared, unambiguous language for developers, evaluators, regulators, and customers to discuss and compare AI system performance. Organizations could better benchmark their systems against industry standards and compare the TEVV results of different models or vendors in a meaningful way. A clear set of standardized metrics could also help catalyze the development of automated tools to calculate these metrics, making rigorous TEVV more scalable and less costly. As part of this effort, AI providers could share their data and test sets to help build a robust catalogue and support fraud and deepfake detection.

Introduce a Clear Severity and Risk Classification Scheme

Simply identifying flaws during TEVV is insufficient; an effective framework must provide a structured way to evaluate their potential impact and prioritize them for remediation. Without a standardized classification scheme, organizations risk misallocating resources by treating a minor issue with the same urgency as a critical vulnerability. Introducing and incorporating a standardized severity and risk classification scheme that would be applicable across model type and use-case would empower organizations to move from a simple list of findings to a prioritized, risk-informed action plan. This could be a matrix that helps organizations triage findings based on their potential impact (e.g., critical, high, medium, low) and

likelihood of occurrence. A formal classification system would enable organizations to respond proportionally and prioritize the most critical risks for immediate mitigation. By including this type of structured, actionable risk classification scheme, the TEVV standard would help provide organizations with a clear, consistent, and defensible methodology for managing the risks associated with AI systems.

Instead of reinventing the wheel, a robust TEVV framework should draw inspiration from well-established and proven risk classification schemes from related fields like cybersecurity and functional safety. Recommending these models provides a common language and a foundation of trust. For example, the Common Vulnerability Scoring System (CVSS) is the de facto industry standard for rating the severity of software vulnerabilities in traditional cybersecurity. Its strength lies in its detailed, quantitative, and transparent approach. The principles of CVSS are directly translatable to AI vulnerabilities. The TEVV framework could propose an "AI-VSS" that adapts these metrics to meet the unique characteristics of AI. This could include expanding the impact metrics to look at security and privacy and expanding the attack vector to include AI-specific vectors such as model interface or user input.

Conclusion

In a fiercely competitive global landscape, the United States' primary advantage will be its ability to deploy AI systems that are demonstrably safe, fair, and secure. Developing an effective TEVV approach that is sufficiently predictive of performance is critical to building the trust in AI systems necessary to deploy and leverage these capabilities at scale and will help ensure that the next generation of AI technologies is built on a bedrock of trust and integrity. TechNet supports NIST’s efforts to broaden participation in and accelerate the creation of TEVV standards that will meet the AI community’s needs and spur greater AI development and deployment. This will require continued collaboration between government, academia, and industry, and we remain eager to partner with the administration in fostering innovation and advancing America’s global AI dominance.

Sincerely,

Linda Moore
President and CEO