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

NOTE TO REVIEWERS

NIST invites input on any aspect of this document, particularly:

- Anything that should be changed about the general direction and structure, including missing sections, material to move or remove, or alternative ways of addressing specific TEVV methods (currently relegated to an appendix given their diversity and, in some cases, lack of maturity)
- The consistency, appropriateness, and sufficiency of the concept map
- How this approach would fit with TEVV practices from sector-specific or non-AI contexts, organizations' approaches to AI TEVV, or existing ISO/IEC guidance on AI testing
- Specific documents (e.g., academic papers or white papers) or other sources of information that NIST should consider or cite
- Specific examples that would be helpful for the appendices
- Any content that is incorrect, not ready for standardization, unclear, or otherwise problematic

Key topics for input are highlighted by call-out boxes in the text.

Input can be shared by email to ai-standards@nist.gov. NIST welcomes input via marked-up documents, bulleted lists of comments and concerns, reaction letters, or any other form that stakeholders find most convenient. Submissions, including attachments and other supporting materials, will become part of the public record and subject to public disclosure. Organizations are also welcome to host listening sessions in which they gather stakeholders to share feedback verbally with the agency.

NIST will consider input received by September 12, 2025 for the initial public draft of the text; input received later will be considered for incorporation into subsequent iterations.

NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Background and Purpose

In March 2025, NIST announced its AI Standards Zero Drafts project. In this pilot project, NIST is collecting input on topics with a science-backed body of work and using it to develop “zero drafts”—preliminary, stakeholder-driven drafts of standards that are as thorough as possible. These drafts will then be submitted into the private sector-led standardization process as proposals for voluntary consensus standards. The project aims to broaden participation in and accelerate the creation of standards, helping standards meet the AI community’s needs and unleash AI innovation.

Based on community input, NIST selected two topics for the Zero Drafts pilot, one of which is AI testing, evaluation, verification and validation (TEVV). This detailed outline proposes a direction and structure for the forthcoming zero draft on TEVV. Based on this paper and input received in response, NIST will propose concrete text for a zero draft, which the community will also be invited to provide input on.

The resulting draft will be submitted to INCITS/AI, the private sector-led committee that represents the United States to ISO/IEC JTC 1/SC 42. SC 42 is the subcommittee that focuses on cross-sectoral AI standards development within ISO and IEC (two prominent international standards developing organizations that collaborate closely through joint committees). Assuming that INCITS/AI proposes the draft as a new project for SC 42 and that SC 42 takes up the proposal, the future of the document will be up to the usual consensus processes of standards development. NIST does not expect to maintain the document further, and will have no greater influence over the resulting standard than a typical contributor to an ISO/IEC national body.

Summary of Approach

NIST proposes to develop a zero draft for a high-level standard on TEVV for AI. Given the rapid pace of development in AI and the extent of flux in AI TEVV methods, the draft is not expected to delve into specific prescriptive recommendations on TEVV methods, nor will it attempt to specify technical aspects in detail. Rather, the proposed standard is meant to serve as an overarching framework that supports AI practitioners in designing appropriate TEVV approaches for specific systems and cases. Relevant factors include, but are not limited to, AI lifecycle stage(s) an organization is operating at and the purposes and objectives it has for conducting TEVV. The framework is envisioned as integrating with current and future ISO/IEC standards on AI testing and providing a suitable backdrop to facilitate their application.

The draft will first aim to establish clear definitions for key terms and concepts in TEVV for AI that will be used throughout the document (mainly to be found in Clause 3), after defining references and scope in clauses 1 and 2. It will then delineate the differences between concepts (e.g., evaluation vs. testing) and the relevant logical requirements for, dependencies between, and limitations of the components of AI TEVV (Clause 4). The draft will focus particularly on when each of these components is applicable and implications of their differences for how AI TEVV activities are conducted (e.g., what kinds of objectives lend themselves to testing vs. other evaluation methods).

TEVV activities are driven by objectives that are typically organizational or socio-technical[1] in nature. Clause 5 will provide a framework that supports practitioners define and explicate relevant objectives, derive appropriate requirements, and then build processes that inform TEVV approaches and methodologies that help determine whether relevant objectives are indeed met. In addition to logical constraints that may in principle prevent addressing certain criteria with certain TEVV approaches, other characteristics may be testable in theory but infeasible to address in practice. To balance needs and limitations, and select the most appropriate approach, organizations need to establish appropriate governance processes related to AI, security, privacy, and information technology and information systems, alongside other relevant areas as they apply to a specific use case, system, or organization.

Two appendices will give insight into the practical application of the framework. The first appendix will provide practical, applied examples of how to perform TEVV for AI systems in practice, demonstrating consideration of the above-mentioned factors. The second appendix will provide a summary of many current technical and socio-technical methods for AI TEVV.

Considerations the Approach Aims to Account For

This proposal is scoped, and its recommendations are designed, to account for several notable challenges in evaluating AI systems:

- AI systems tend to exhibit multiple levels of complexity that interfere with TEVV. Components like models and datasets are often inherently complex and are often difficult to decompose into easily evaluated units. They are also embedded in varied technical architectures, organizational configurations, and contexts of interaction with humans.
- These issues influence how TEVV can be conducted; e.g., formal verification of large models is not generally achievable, and evaluation results will often reveal tendencies or likelihoods rather than definite measures.
- With AI systems that are deployed to deal with high-level tasks, objectives and requirements are often difficult to appropriately operationalize and measure, similar to challenges found in other areas that deal with complex concepts and systems.
- Many different entities and organizations may be performing TEVV for many different reasons (e.g., to determine market readiness, for conformity assessment, or to assess whether a product meets a specific need).

[1] i.e., involving interactions between technical systems and people.

- The quality, accuracy, coverage, and definiteness of evaluations and their results need to be contextualized and understood correctly, with appropriately managed expectations. For example, recipients of evaluation results will need to recognize that many findings will only be probabilistic and not provide certainty. Setting expectations via communication and documentation is especially critical when evaluation results are supplied to other parties.

Proposed Document Structure

Since NIST aims to propose this standard for development in ISO/IEC, the below structure is based on conventional ISO/IEC structures and generally retains a similar approach to ISO/IEC 42001 (Information technology — Artificial intelligence — Management system) and ISO/IEC/IEEE 29119-1 (Software and systems engineering — Software testing — Part 1: General concepts) where feasible.

Clause 1: Scope

Per ISO/IEC convention, this clause will consist of a terse paragraph or two specifying what the rest of the document will cover. It will clarify that the standard will circumscribe key terminology, concepts, processes, and fundamental requirements of TEVV approaches and methods independent of sector and AI system type. It will also emphasize governance processes given the socio-technical nature of AI TEVV.

Given the complexity, variation, and context-sensitivity that are characteristic of AI systems in situ, this standard can only provide a framework and toolkit for those conducting TEVV. The document will not aim to be specific or technical in nature; the field is developing too rapidly to allow for such standardization at this point.

Clause 2: Normative References

Normative references will be added as appropriate to incorporate definitions or requirements external. References in Clause 2 are limited to existing ISO documents.

Clause 3: Terminology

Per the usual ISO/IEC document structure, Clause 3 will contain definitions of key terms that underlie the AI TEVV concepts and framework.

These terms will include at least the following:

- Testing
- Evaluation
- Verification
- Validation

- Validity (of a TEVV approach)
- Reliability (of a TEVV approach)
    - Possibly: Specific types of reliability and validity, as they relate to TEVV
    - Measurement error

Additional terms, such as terms for specific methods, will be defined as needed for the body of the document.

Definitions will be crafted to reflect the conceptual discussion in Clause 4, and will be aligned as closely as possible with existing ISO/IEC definitions. Elaboration on relevant issues for practitioners, such as validity, reliability, sampling, and practical feasibility, will be left for Clause 4, as required by ISO directives on terminology.

Clause 4: Key Terms and Concepts in TEVV for AI

This clause will provide a conceptual overview of what TEVV for AI consists of, starting with how the key concepts relate to each other. The aim is to establish clear concept maps; hierarchies that indicate superordinate and subordinate concepts, where appropriate; and an account of AI TEVV's notable requirements and limitations.

Some key terms are used in ways that are inconsistent with standard usage in other relevant fields, or even inconsistently within AI. While this is not always problematic, the framework in this zero draft will aim to be clear, usable, and logically consistent, both internally and with respect to non-AI-specific TEVV activities:

- The framework will provide a foundation for consistent discussion via a system of terms and concepts that can be used uniformly within AI to describe AI TEVV activities and find agreement on effective approaches. For example, a shared understanding and approach to verifying a system would simplify communications between stakeholders, internal decision-making, and third-party supply chain management.

- The framework will also aim to maintain maximal consistency with TEVV terminology and concepts from domains outside of AI. For example, some common uses of “red teaming” in AI are at odds not just with other uses within AI but also with other fields. The zero draft will seek to avoid such inconsistencies. Remaining consistent with other fields—and if possible general understanding—is increasingly important as AI moves into general use, and AI TEVV outcomes are thus integrated into relevant reports, audit proceedings, system testing, etc.

> As TEVV-related activities are often associated with quality, safety, and other regulations, practices, and standards in other fields (e.g., medical care, manufacturing, and nuclear energy), the draft will track relevant intersections.

However, there are some inconsistent uses of similar terms that are unproblematic. For example, “test data” or “test time” in AI often refers to running a system on non-training inputs, whereas this document is concerned with “testing” as an evaluation activity. The similarity in terminology does not present a problem so long as uses of the two concepts are distinguishable. This document will focus on terms as they relate to TEVV and assurance activities in general, even when these terms have AI-specific meanings unrelated to TEVV.

Outline:

- Common characteristics among and relationships between TEVV concepts
    - TEVV in general speaks to ascertaining the extent to which a target—e.g., a product or service—meets defined objectives, or helping others determine if their requirements will likely be met. This could include assessing features, qualities, performance, and other characteristics.
    - There are significant differences between testing, evaluation, verification, validation, and related concepts, in terms of both the meanings of the terms and what is required to execute the work they describe (see Table 1 and Figure 1 for NIST's current conceptualization of the terms and their relationships).
    - Individual methods cited in Figure 1 can be combined into an overarching evaluation approach. For example, an evaluation approach would be created to address TEVV requirements and objectives case-by-case and may then include starting with user research using interview methods followed by benchmarking against user needs.

> NIST particularly welcomes input on this conceptual material, which NIST aims to expand with relevant practical dependencies and procedural links between types of and methods for TEVV. In the zero draft, most methods in Figure 1 will be elaborated upon in detail only in an appendix, if at all.

Term	Relationship to other concepts	Current understanding	Examples
Evaluation	Superordinate concept that can be specified further, and that can be implemented via a variety of methods and approaches.	Quantitative and/or qualitative determination if or to what extent an evaluation target meets or exceeds a set of characteristics or criteria, based on an appropriately defined and documented set of items (objectives, requirements, specifications, metrics, measurements), using a defined, clear and documented methodology.	Organization evaluates different software vendors to find the best one for its needs; conformity assessors evaluate whether an AI system conforms to a standard.
Verification	Type of evaluation with a particular goal	Verification activities are conducted to ensure that the design and development outputs meet the input requirements (from ISO 9000), i.e., the specifications it is being constructed to.	Software provider seeks verification that an AI system they have developed meets the specifications they received from the client.
Validation	Type of evaluation with a particular goal	Validation activities are conducted to ensure that the resulting products and services meet the requirements for the specified application or intended use (from ISO 9000). Validation requires that the use case, context, and the specific requirements of the organization are known; unlike for verification, this requires insight into the user of a system (or the "system owner").	An organization procuring an AI system from a provider validates that the system meets its business needs.
Testing	One methodology among many for evaluation	Strictly specified determination of one or more measurable and operationalized characteristics of an object of assessment, according to a specified, repeatable, measurable, achievable, and relevant procedure. A procedure is a specified way to carry out an activity or a process. (Based on ISO 17000 family)	An AI system is scored on how many inputs from a specified dataset it gives canonical responses for; an AI system is presented with many queries at once to assess its performance under load.

Table 1

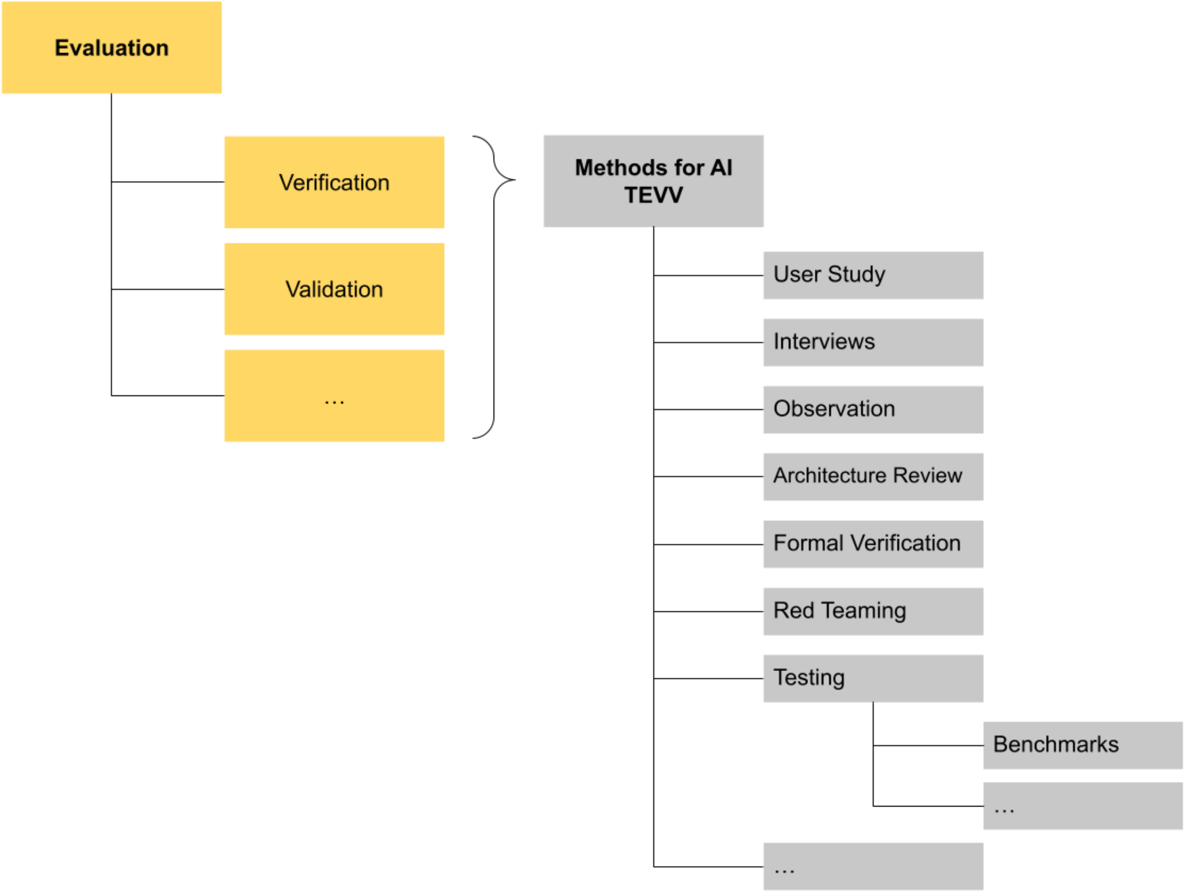

Figure 1

- Considerations for each of testing, evaluation, verification, and validation:
    - Requirements
        - E.g., data availability, information about the system owner and/or use case, etc.
    - Validity
    - Reliability:
        - Particular focus will be on different conceptions and types of validity and reliability.
    - Sampling of data and cases
    - Selection of approaches and methods
        - E.g., when statistical and quasi-experimental methods are helpful
    - Practical feasibility

> NIST particularly welcomes input on specific considerations that should be included for testing, evaluation, verification, or validation.

- Limitations of TEVV, particularly TEVV for AI, and how to address them in practice
    - Many AI systems cannot be easily tested against some likely objectives because relevant factors and variables are difficult to specify and/or keep constant.
    - Some characteristics or objectives are not amenable to evaluation at the level of rigor that would be required for high assurance engagements.
        - For example, some requirements or objectives cannot be operationalized sufficiently for verification, or turned into testable criteria.
    - Evaluation results are generally considered to be time-bound.
        - This limitation is particularly pronounced for AI systems: the environments they operate in can change over time, which can alter the statistics models rely on, and some AI systems evolve over time as well (e.g., through continuous learning).
    - Contemporary AI systems exhibit complexity at multiple levels, similar to other systems but with particularities that add to or exacerbate the complexity. Issues at each level propagate to the higher levels as well.
        1. Individually complex components
            - AI systems often have at their core intrinsically complex components whose behavior cannot readily be analyzed.
                - For example, the complexity of large machine learning models hinders understanding and thus the explainability and traceability of individual outputs.
            - AI models are typically built with large amounts of training data. These training data may not be possible to evaluate individually.
                - E.g., the training dataset may be fully or partly unavailable or simply too extensive to review.
        2. Complex components that are arranged into complex technology stacks
            - Complexity in technology stacks can often be managed by decomposing the system into sub-units, evaluating them individually, and evaluating their integration. This is challenging for many AI systems for several reasons:
                - AI models often exhibit unpredictable behavior, making it hard to reason about how they will influence other parts of a technology stack built on top of them.
                - The training data can influence the behavior of models built on them in inscrutable ways, including via

interactions between training samples that may be hard to reason about.

- Most AI users rely on large, complex, distributed supply chains involving, for example, varied technical components and a variety of organizational or contractual configurations. This can inhibit TEVV, for example by limiting individual actors’ ability to readily understand components created by others.
    - For example, a model developer would have deep insights into the model and its development but they are unlikely to be privy to how that model would be used by resellers, service providers, and the end user.
    - The end user, on the other hand, would be aware of the specific requirements and context, but would lack developers’ technical insight.

3. Complexity that emerges from interactions between components and human behaviors or between components with many-directional influences between them.
    - AI system architectures and systems often cannot be conceptualized and evaluated as a stack with neat hierarchies and abstraction barriers. Instead, many AI systems rely on technical components that interact with each other within intricate networks, further limiting the possibilities to decompose and simplify by evaluating components individually.
    - The complexity is further exacerbated by AI systems’ deployment context. Individual components and the system as a whole interact not just with their technical context, but their socio-technical context. Particularly given AI systems’ unpredictable behavior and hazy abstraction barriers, components or systems that appear to work well in one evaluation context may not perform similarly in the deployment context where humans or organizations interact with them differently.

- These complexities yield practical challenges for evaluators, including:
    - Evaluation methods based on decomposing a system (e.g., unit and integration testing) are often difficult to apply or need to be adapted.
    - For many AI systems, such as those based on large language models, certain characteristics are likely impossible to ascertain with complete certainty.

- Evaluators need to ascertain which systems and characteristics can be tested, evaluated, verified, and validated, and to what extent.
- Higher complexity yields a combinatorial explosion of system elements and situations that one might want to evaluate—but organizations have finite resources and time for evaluation.
  - For example, testing complex systems requires more tests and test cases to reach valid and significant conclusions. However, there are practical limits on the number of tests and test cases.
- Rather than producing highly reliable results, evaluators will often have to rely on probabilistic findings.
  - Such findings require more careful crafting and interpretation to yield useful insights.
  - This shift also introduces new methodological issues, such as ensuring that probabilistic findings are based on samples that allow for generalization.

- Ways of handling these limitations:
  - Many experts draw from methods inspired by fields that regularly grapple with complex human behavior and confounding variables, such as psychology, anthropology, and other social sciences.
    - Many such methods come with their own requirements and usually decreased precision compared to TEVV methods for more straightforward targets. For example, while qualitative methods such as user interviews can be very useful, they may introduce greater subjectivity if not conducted with care.
    - While there are similarities between some social science methods and AI TEVV, the needs, concepts, and principles of these fields are not always the same. Thus, relevant social science knowledge and expertise must be adapted appropriately.
  - When facing characteristics or objectives that cannot be directly evaluated, it is often possible and worthwhile to evaluate systems based on abstract concepts that are not fully operationalized.
    - E.g., user satisfaction could be operationalized into more granular metrics amenable to automated testing, but it could also be left abstract and assessed through user surveys.
    - Such an approach may come at the cost of narrowing the range of methods and approaches that can be used, reducing validity, precision, and reliability, and increasing expense, duration, and complexity.

            - Reducing the specificity or “richness” of the concepts used in evaluation may lead to better reliability but also make the evaluation less detailed and comprehensive in terms of what can be learned about the target, in turn compromising validity.
        - Many challenges can be mitigated by well-planned, extensive TEVV procedures with clearly defined steps, resources, inputs, outputs, etc.
            - In particular, it is key to keep track of, and communicate appropriately, what findings and results mean in context of system, architecture, objectives and environment. This includes communicating the probabilistic or imprecise nature of many findings.
            - Depending on the objectives or requirements of the system or the evaluation, TEVV plans may need to include continuous evaluation or regular re-evaluation to ascertain the validity and reliability of results over time.
- Documentation needs and requirements
    - Documentation can help ensure that TEVV engagements are appropriately communicated. A key element of many TEVV engagements is to build trust or otherwise provide evidence.
    - Thus, among other relevant aspects, documentation should clarify:
        - How the engagement is based on appropriately defined and measurable objectives that were determined in a functional, appropriate, and repeatable manner with suitable input from interested parties.
        - How characteristics, data points, measures, and other relevant factors or variables have been appropriately defined, operationalized, and measured.
        - How the engagement followed appropriate procedures throughout.
        - How the evaluation results’ applicability can be expected to change over time and under various environmental shifts.
    - Due to the above-mentioned complexity of systems and objectives in AI evaluation, methodology sections should be more detailed and contain more discussion about choices and rationale than may be needed in other areas. Critically, readers should be given the information necessary to interpret the report in a consistent manner, including on the following topics:
        - What could and could not be measured
        - How accurate, generalizable, and uncertain results are
        - What environmental or contextual factors are pertinent to the outcome and results
        - Why the evaluators made certain choices, e.g., to prioritize one evaluation objective over another

Clause 5: Governance, Process, and Organizational Requirements

This clause will provide an account of what needs to be in place for an organization to conduct TEVV successfully. To appropriately test, evaluate, verify, or validate a system, organizations need to establish system objectives and characteristics, define how best to operationalize them, and establish processes and measurements that are consistent, functional, and repeatable such that they can consistently yield reliable results over multiple assessments and include the necessary input.

As outlined in figure 2, organizational requirements, as well as other factors, determine or inform the specific objectives and requirements of an AI evaluation engagement, which in turn drive the evaluation activities. The objectives, requirements, and evaluation activities are additionally constrained by a variety of factors such as technical limitations and budget. Furthermore, ontological and epistemological questions, particularly with respect to validity, reliability, and case sampling, should be considered when settling on objectives, requirements, and evaluation activities and methods. Though such concerns may seem abstract, they have very practical implications for whether a given methodological approach to TEVV will achieve usable outcomes. For example, if an organization were to incorrectly operationalize the objective of system security to be limited to the model instead of the AI system as a whole, the findings would not actually reflect system security, as they would exclude relevant areas such as infrastructure.

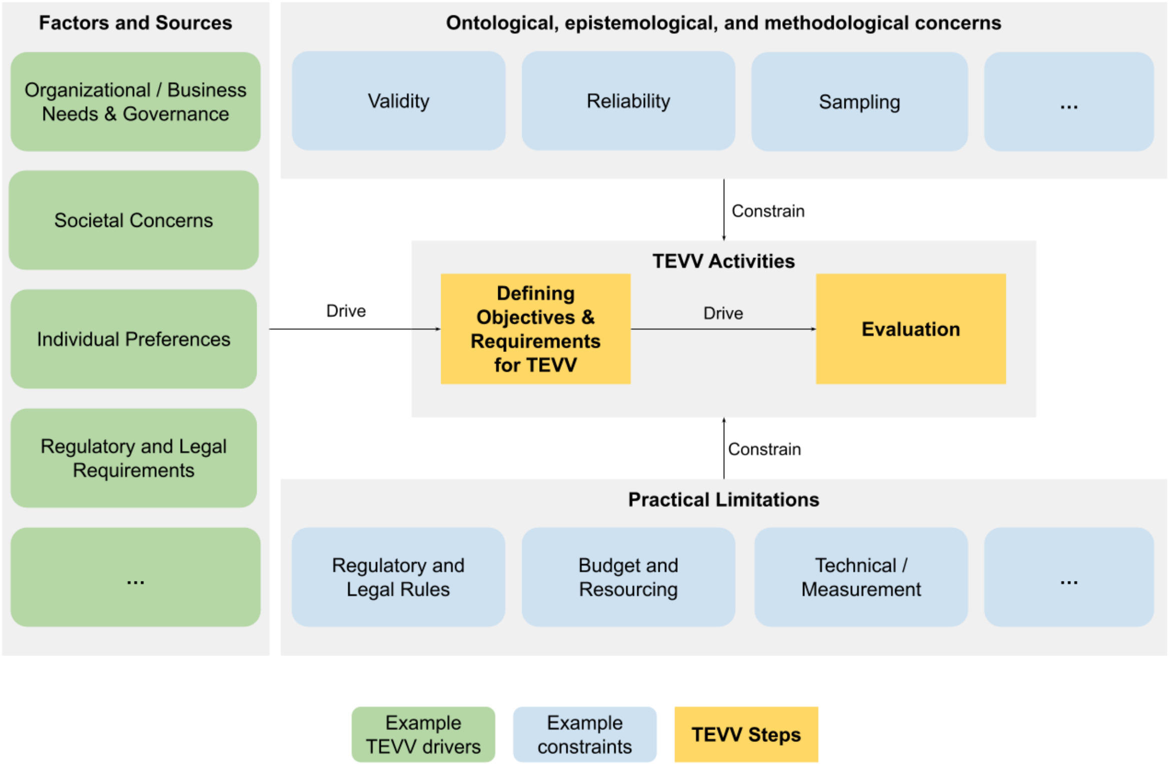

Figure 2

Outline:

- Guidance on how to establish suitable requirements with input from interested parties. These include business, system, or process-related objectives.
    - Typical requirements or concerns that need to be addressed, for example in the areas of:
        - Security
        - Functional safety (e.g., prevention of injuries)
        - Prevention of system abuse or misuse (i.e., use of systems outside of their intended usage to deliberately cause harm)
        - Privacy
        - Quality
        - Performance and efficiency
    - Which requirements, qualities, or aspects can or cannot be reasonably assessed, and how. As discussed above, complex objectives can be more difficult to operationalize into criteria that are straightforward to assess, while some high-level objectives or interests in the AI community, especially heavily debated issues such as some related to ethical behavior, may never be fully addressable by TEVV methods.
    - The effectiveness and efficiency of different approaches vis-a-vis organizational, technical, and intrinsic methodological constraints.
    - Ensuring the appropriate resourcing, support, and independence of the assessors.
    - Continual verification and updating of the established objectives, criteria and variables.

- Guidance on establishing, operating, reviewing, and continually improving support processes that surround and support actual testing, feeding into the assessment, or providing responses and follow-up to assessments. Processes and support should be maintained for:
    - Business needs, strategy, governance
    - Process definition, creation, execution
    - Budget and support
    - TEVV requirement setting
    - Engineering and technical requirements
    - Execution
    - Documentation
    - Response
    - Interested parties and their needs, e.g. users and consumers

- Guidance on navigating and accommodating limitations and constraints in AI assessment will be a common challenge. This subclause will thus provide specific guidance on how to accommodate different organizations, budget levels, areas, and sectors, elaborating on how to balance assessment effectiveness with limited resources.
    - Analysis, assessment, and decision-making on trade-offs
    - Risk-based optimization of assessment processes
    - How organizations can address typical limitations:
        - Technical limitations
        - Intrinsic methodological limitations
        - Temporal limitations
        - Budgetary constraints
        - Skill and ability constraints
        - How to manage third parties and supply chains

> NIST particularly welcomes input on specific pieces of guidance the zero draft can offer on each of the sub-bullets in this clause.

Appendix 1: Examples of performing TEVV

Fundamentals, Governance, Operations, and Support

The first appendix will provide more applied examples of how the content in clauses 5 and 6 can be applied in different practical scenarios, given specific systems, aims, and constraints.

This appendix will outline how to address the following aspects of preparing for TEVV through the worked examples:

- Governance and Strategy
- Requirement setting
- Process definition, creation, execution
- Documentation, Response and Improvement

The following items are expected to be key variables for elaborating on how to make decisions and build assessment approaches in practice:

- System use case and context
- Organizational requirements
- The legal and regulatory environment
- The position of the assessing organization in the AI supply chain
- The maturity of the system and its current status in the AI systems lifecycle

TEVV Process

Based on established objectives and criteria, the appendix will describe how to operationalize the output of governance processes and conduct TEVV. Each example case will be discussed as an application of the aspects below:

- Use of well-known approaches and structures, such as the V-model, in the context of AI
- Engineering and technical requirements
- Operationalization of objectives
    - For example, in keeping with the NIST AI Risk Management Framework, concepts like valid/reliable, explainable/interpretable would have to be operationalized into variables that can be measured.
- Measurement considerations, including ontological, epistemological, and methodological concerns
- TEVV execution
    - Summarizing different TEVV strategies and technical methods, and what outputs they can and cannot produce in the example cases
        - E.g. this may depend on required inputs, the situation and context, etc.
        - Limitations, strengths and weaknesses, and contamination avoidance as applied to specific cases
- Appropriate TEVV procedures, compliance, and documentation
- Reporting and communications
- How to respond to findings, as it relates to TEVV execution (the organizational response is covered above)

Appendix 2: Technical and sociotechnical approaches and methods

While some areas of AI, such as supervised machine learning, are relatively mature and offer well-established TEVV methods, AI TEVV methods and approaches overall, especially for in-situ evaluations, are in considerable flux. Acknowledging this limitation, Appendix 2 aims to provide a necessarily incomplete catalog of commonly used methods, approaches, and techniques.

Each item, will include the following information:

- Description of the approach
- Required inputs and expected outputs
- Situation where this approach may apply
- Characteristics of the approach if applied in practice
    - Limitations
    - Strengths and weaknesses
        - E.g., contamination issues

    - Which topic areas or questions the approach is suitable for; this may include discussions of fit, e.g., at which life cycle stages an approach or method may be most beneficial.
- Relevant references
- Pointers to other items that are relevant, e.g., linking higher-level approaches to specific techniques that integrate well with each other