Measuring and Influencing Trustworthiness of AI Enabled Systems

September 2025

Carol Pomales
Elena Charnetzki
MITRE Support
Developmental Test, Evaluation, and Assessments

Controlled by: D(DTE&A), OUSD(R&E)
CUI Category: n/a
Distribution: Distribution A – Unlimited; DOPSAR Case 25-T-3150
POC: Mr. Orlando Flores, 571.3724145

BLUF

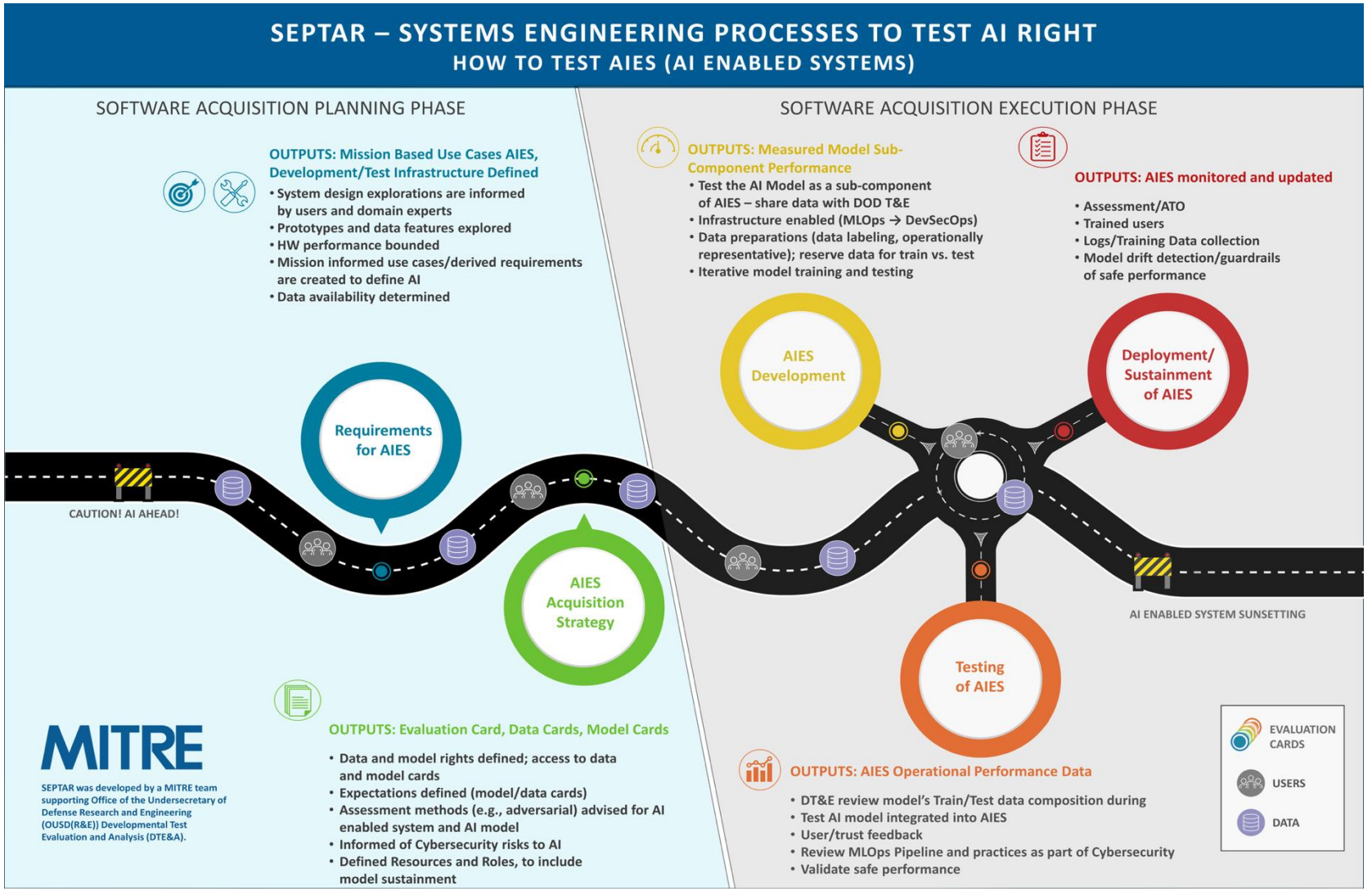
- AI-enabled systems (AIES) in the Department of Defense (DoD) will benefit from shifting our thinking of T&E as single event towards a process that takes proactive measures from the start of the system lifecycle and produce evidence to inform risk
- To accomplish this, we propose a trustworthiness metric to inform project and test teams on how to make choices and enable key activities that mitigate risks
- Our definition of trustworthiness is framed using research on emerging best practices for designing, building, and employing AIES from DoD practitioners, academia, federally-funded research & development centers (FFRDCs), and commercial industry
- By measuring trustworthiness, the tester can apply a repeatable process to use insights into the development process to plan for the necessary testing for the AIES

Overview

1. Background
2. Definitions
3. Approach
4. Benefits
5. Examples
6. Questions

Motivation/Background

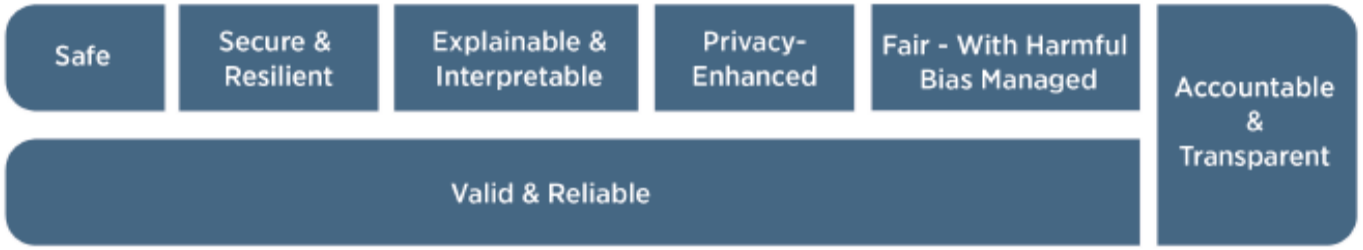
- Work performed in support of the Developmental Test, Evaluation, and Assessments (DTE&A)
- Test and evaluation (T&E) of AIES will be a complex challenge for the DoD T&E
  - Assuring and understanding the processes used to build the AIES informs on the later T&E
  - Through ongoing collaborations, we have defined actionable recommendations that can be measured.
- We seek to provide a <u>trustworthiness metric</u> to facilitate the approach to T&E AIES



Trustworthiness Defined

Assessing Trustworthiness of the AIES is useful to the tester to help define their approach to T&E and for the broader team to ensure a more adaptable and efficient approach

- NIST has provided useful common framework informing the many dimensions of Trustworthiness
- Definitions for Trust and Trustworthiness are often confused or misquoted; NIST has provided definitions for Trust and Trustworthiness that DOD now follows



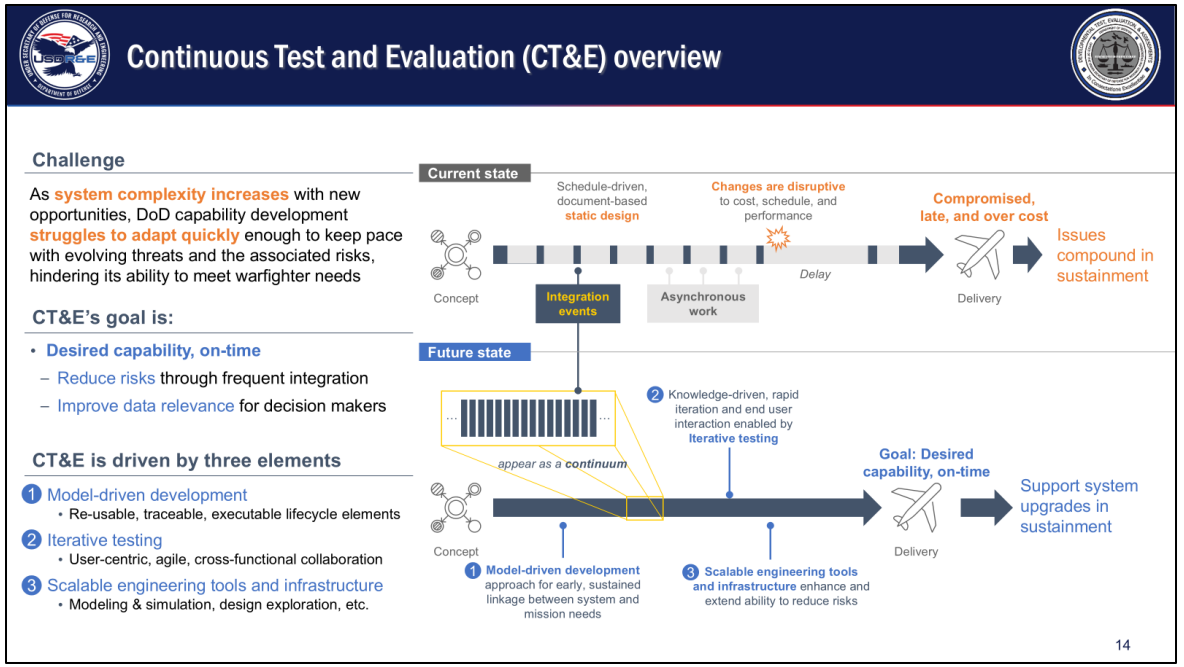
Characteristics of Trustworthy AI Systems [1]

Trustworthiness = The degree to which an information system (including the information technology components that are used to build the system) can be expected to preserve the confidentiality, integrity, and availability of the information being processed, stored, or transmitted by the system across the full range of threats and individuals' privacy [2]

Trust = the system status in the mind of human beings based on their perception of and experience with the system; concerns the attitude that a person or technology will help achieve specific goals in a situation characterized by uncertainty and vulnerability [2]

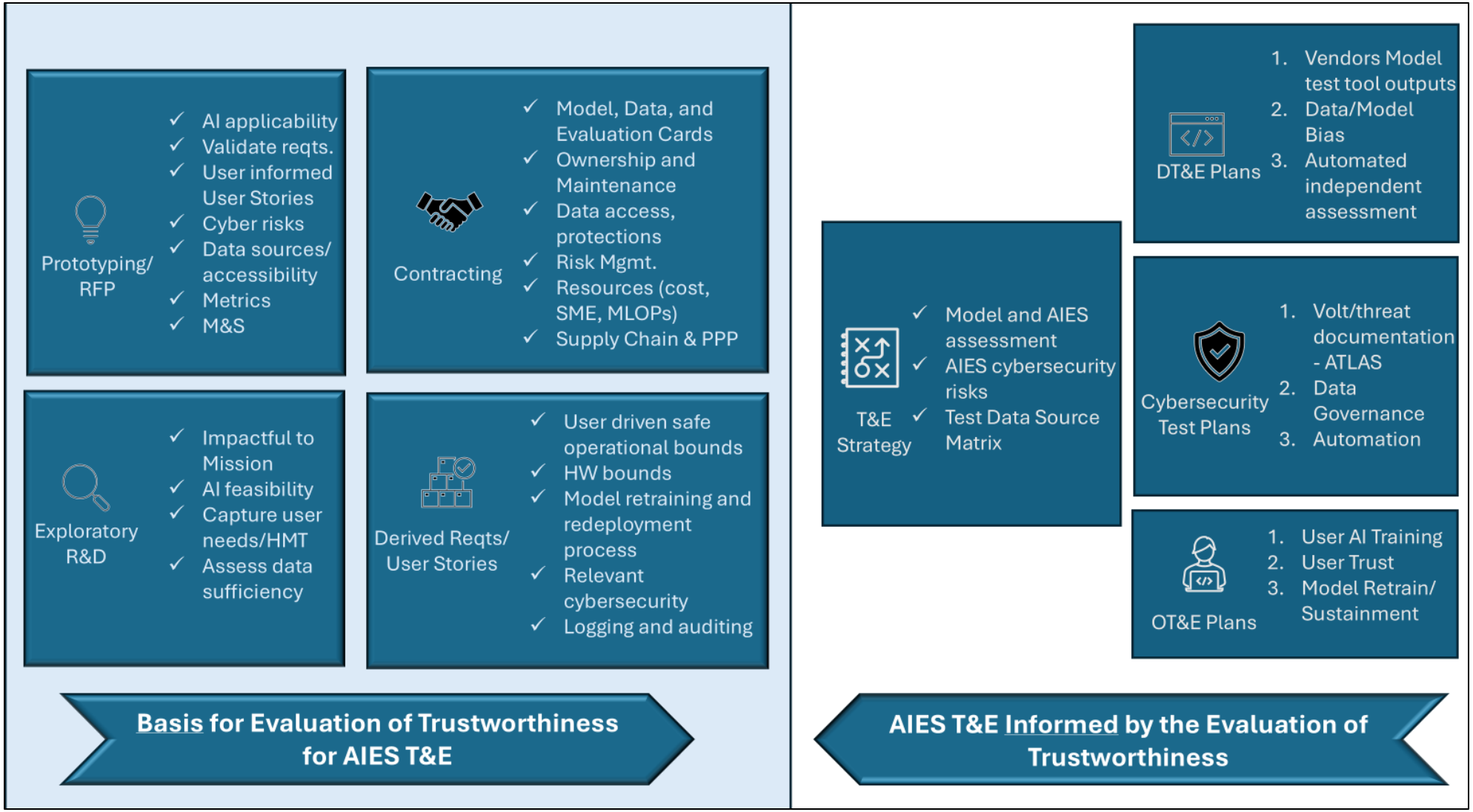
Trustworthiness Metric Defined

- Trustworthiness is a multidimensional metric informed by the activities and decisions made across the AIES lifecycle aligned to the principles of CT&E
- The proposed **trustworthiness metric** to quantitatively measures the choices and key activities that produce an AIES and informs action
- Measurement is framed upon research of emerging best practices for designing, building, and employing AIES from DoD practitioners (e.g., CDAO, TRMC, DOT&E, Service PMOs), academia, federally funded research and development centers (FFRDCs), and commercial industry
- The trustworthiness metric provides insight into a more streamlined AIES development process to inform scope of testing
- The output is a set of recommendations that will help to inform the T&E community as they make decisions on how to more effectively and efficiently plan to evaluate the AIES



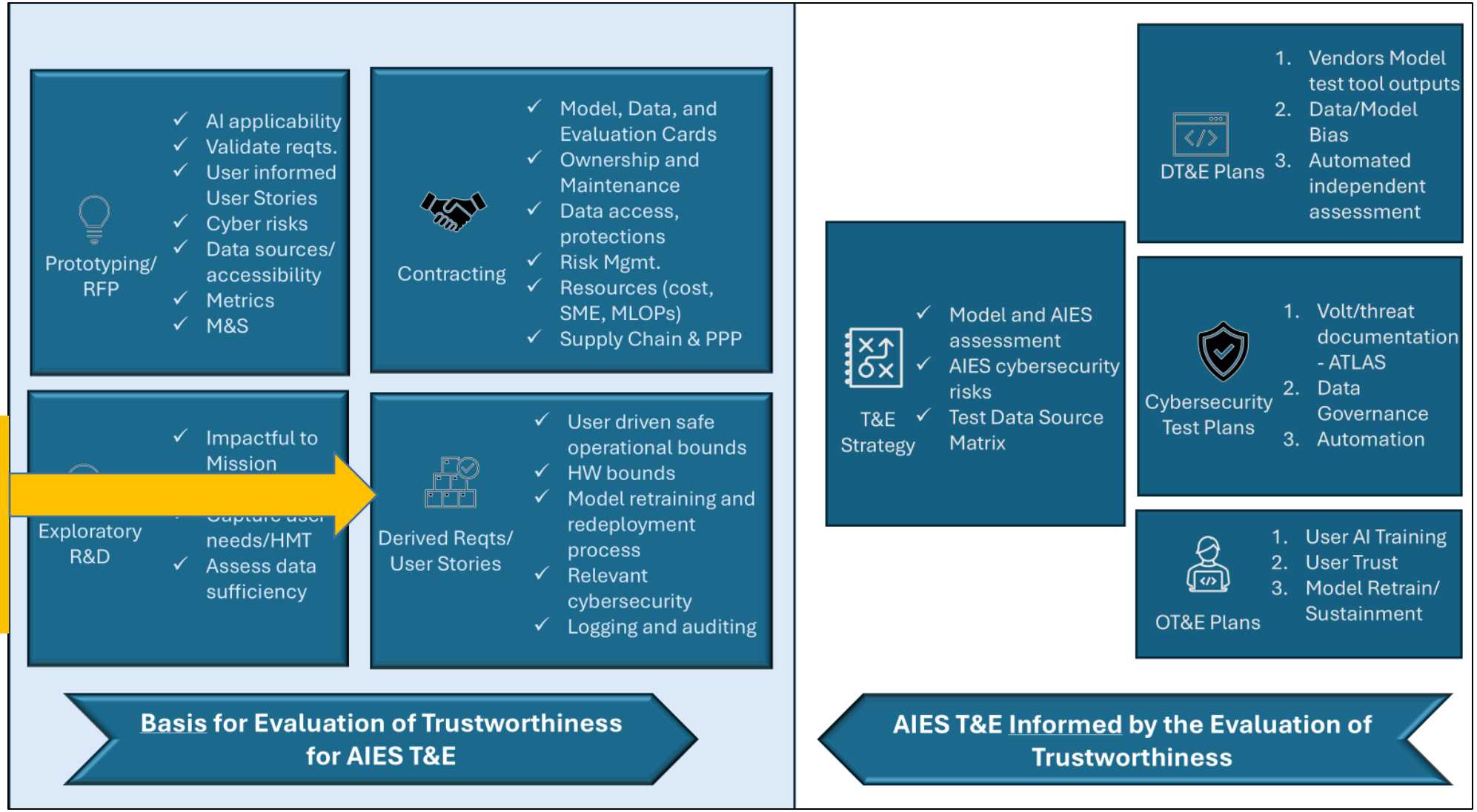
Trustworthiness Approach

- This is a top-level view of the key activities that are assessed by the trustworthiness metric.
- Each assessment comes with complementary T&E recommendations.
- This metric can be measured at various times, but last assessment needs to occur prior to the T&E Strategy.



Trustworthiness Approach

We will further explain one of the dimensions, 'Derived Requirements', on the following slide.



Derived Requirements - User Stories Checklist

- Top level requirements are decomposed by program offices into derived requirements/ User Stories to drive development of the AI components
- The derived requirements check list assesses their rigor and how they will increase the trustworthiness of the system

☑	Are developed with the involvement of the AIES components' end users
☑	Specify APIs/protocols for secure data sharing and data transfers between internal and external systems
☑	Prioritize the development of interpretable and explainable AI models
☑	Identify potential (model/data) biases up front and specify bias mitigation methods
☑	Specify capabilities for activity logging and system auditing AI model performance
☑	Define system operations beyond unexpected or safe operating bounds of the AI model
☑	Specify procedures for monitoring model performance to inform when model retraining should occur.
☐	Define hardware performance boundaries for the deployed AIES.
☐	Specify capabilities/procedures to swap out model for retraining
☑	Specify methods and procedures for users to provide feedback on model recommendations, provide corrective model feedback, and to report other system issues.
☑	Specify cybersecurity mitigation capabilities features in the system's design specifications that are based on VOLT (or threat documentation), and may include access controls, firewalls, intrusion detection systems, and strategies for incident response.
☑	Detail specific capabilities that are built as part of the inherent system that enable the deployment and sustainment of the AIES.
☑	Include privacy protection measures that are based on data protection laws and regulations, such as NIST Privacy Framework or the Defense Privacy and Civil Liberties Division.
☑	Are updated as the AI system design warrants evolution.

Derived Requirements - Trustworthiness Checklist Approach

- The 'Derived Requirements' checklist is performed by the test lead in collaboration with the PM Leader and other SMEs
- After completing the trustworthiness metric's checklists, the tester can review the resulting recommendations to inform the T&E Strategy and Test Plans
- Iterative evaluations of trustworthiness early in the lifecycle can allow for corrections and improvement to occur, benefiting the program
- The benefit of this metric evaluation to the tester is the creation of a more repeatable process – growing a knowledge base to help with this challenging technology
- PMs can use this information to build in capabilities to inform on performance and also gain an early understanding of the implications of AI technology to subsequent T&E complexity and risk

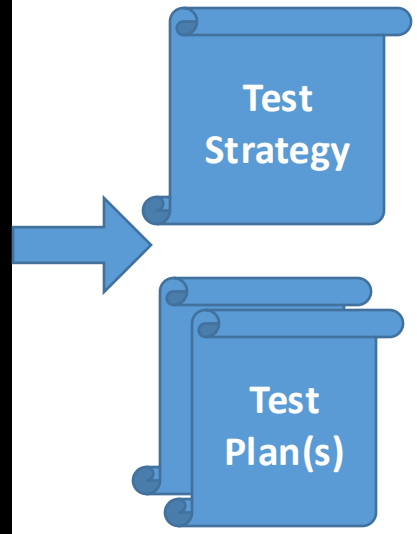
Example Checklist Output & Recommendations for Test Strategy and Test Plans

Hardware bounds defined for the AIES?

- *Yes-> plans should test to defined bounds and note implications to use and sustainment.*
- *No ->, test to failure (HW) and note implications to use and sustainment.*

Capabilities and Procedures to Swap the Model?

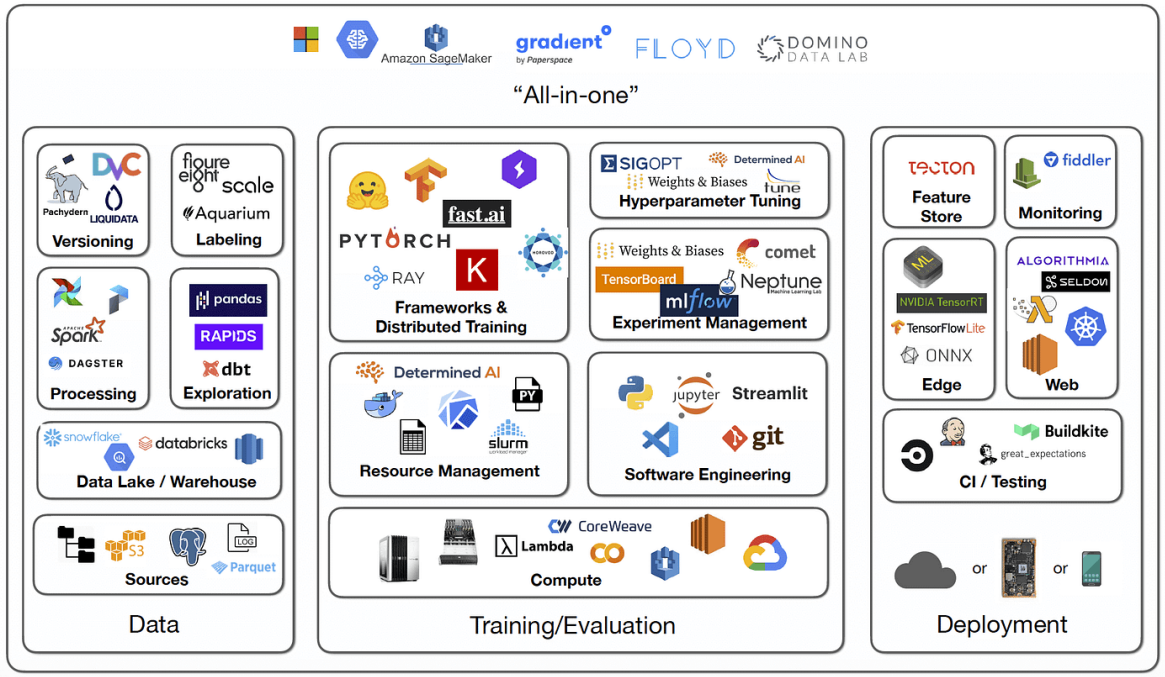
- *If yes, explore the documented procedures during evaluation by an AI SME*
- *If no, request a demonstration of the process and infrastructure for the AI model(s) to be retrained and redeployed during live testing and note any concerns or risks to mission (e.g., OPTEMPO)*



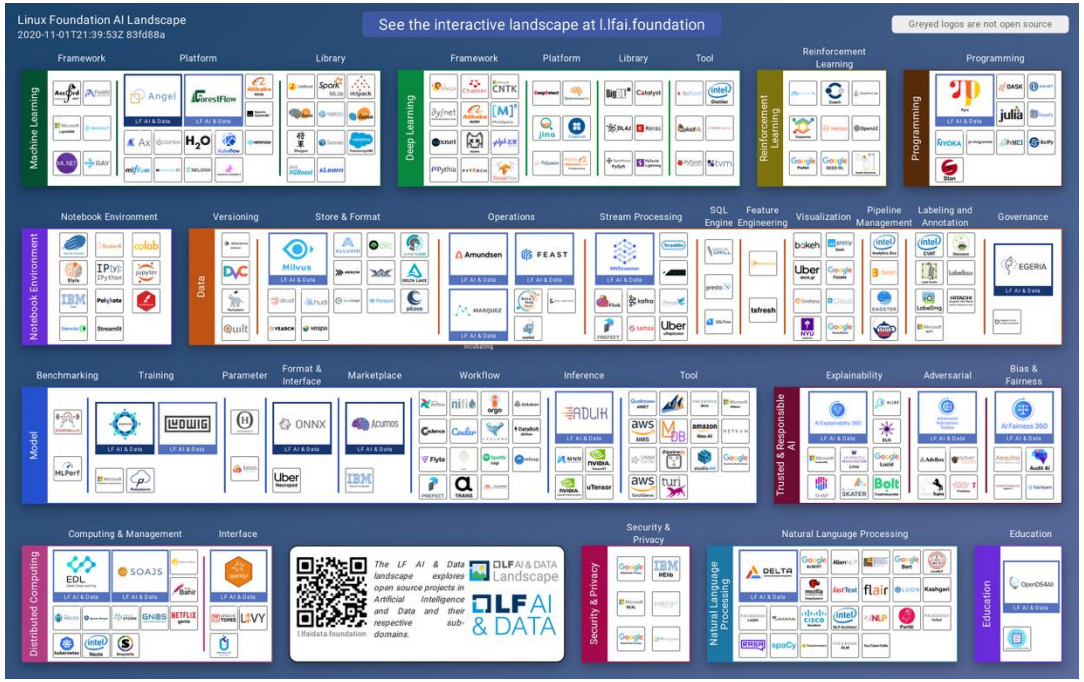
Tools

- Seek to have commercial or DoD Tools (e.g., CDAO, TRMC), automate all or portions of the trustworthiness assessment

- Need to have programs and industry move towards a future of automation of T&E of AIES, where information can be collected in a trusted manner that facilitates an efficient T&E planning approach



https://medium.com/aiguys/mlops-development-infrastructure-and-tooling-acb53b5fe28e



https://ml-ops.org/content/state-of-mlops https://landscape.lfai.foundation/

Advantages

- Help test and evaluate AIES more quickly
- Provide information to refer to as testers are engaging earlier in the process
- Assist testers who are new to AI technology
- Provide common and deliberate practices and techniques to perform testing based on broader AI SME and insights
- Inform leaders on the risks and rewards of using AIES

Next Steps:

- Continue to map Trustworthiness to the MLOps process
- We seek collaboration partners to further validate these recommendations through pilots; these insights will help to inform DoD Guidance and Policy for T&E of AIES

MITRE POCs:

Carol Pomales – cpomales@mitre.org
Dr. Natalie Kautz – nkautz@mitre.org

References

[1] National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” U.S. Department of Commerce, Washington, D.C., 2023.

[2] National Institute of Standards and Technology, “The Language of Trustworthy AI: An In-Depth Glossary of Terms,” U.S. Department of Commerce, Washington, D.C., 2023

[3] C. Balhana, I. Chen, R.W. Ferguson, J. Lockett, D. Moore, C. Pomales and F. Reeder, “Systems Engineering Processes to Test AI Right,” The MITRE Corporation, McLean, 2023.

[4] C. Collins and K. Senechal, “Test and Evaluation as a Continuum,” *The ITEA Journal of Test and Evaluation*, vol. 44, no. 1, 2023.

Continuous Test and Evaluation (CT&E) overview

Challenge

As **system complexity increases** with new opportunities, DoD capability development **struggles to adapt quickly** enough to keep pace with evolving threats and the associated risks, hindering its ability to meet warfighter needs

CT&E's goal is:

- **Desired capability, on-time**
  - Reduce risks through frequent integration
  - Improve data relevance for decision makers

CT&E is driven by three elements

1. Model-driven development
   - Re-usable, traceable, executable lifecycle elements
2. Iterative testing
   - User-centric, agile, cross-functional collaboration
3. Scalable engineering tools and infrastructure
   - Modeling & simulation, design exploration, etc.