Early and Developmental Test and Evaluation of AI-Enabled Systems

Sara Jordan & David Sparrow

Developmental Test and Evaluation of Artificial Intelligence-Enabled Systems Guidebook

Office of the Director, Developmental Test, Evaluation, and Assessments
Office of the Under Secretary of Defense for Research and Engineering
3030 Defense Pentagon
Washington, DC 20301
osd.r-e.comm@mail.mil
https://www.cto.mil/dtea/

June 30, 2025

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

Expectations for this Webinar

What you can expect

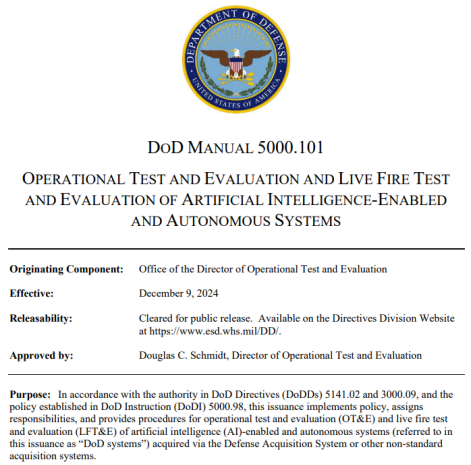
- Background for this guidebook
  - Associated guidebooks
  - The need for this guidebook
- Content of the Guidebook
  - How AI might change T&E
  - AI Development
  - AI data and VV&A
  - Synthetic data
  - Model visibility
  - AI and the CONEMP

Next steps for you

- Download a copy of the guidebook
- Participate in the Q&A
- Provide feedback
- Express your interest to participate in version 2

Context of the Guidebook

Current T&E Policy and Guidance is Evolving to meet the AI moment



DoD Manual 5000.101

Operational Test and Evaluation and Live Fire Test and Evaluation of Artificial Intelligence-Enabled and Autonomous Systems

Originating Component:	Office of the Director of Operational Test and Evaluation
Effective:	December 9, 2024
Releasability:	Cleared for public release. Available on the Directives Division Website at https://www.esd.whs.mil/DD/.
Approved by:	Douglas C. Schmidt, Director of Operational Test and Evaluation

Purpose: In accordance with the authority in DoD Directives (DoDDs) 5141.02 and 3000.09, and the policy established in DoD Instruction (DoDI) 5000.98, this issuance implements policy, assigns responsibilities, and provides procedures for operational test and evaluation (OT&E) and live fire test and evaluation (LFT&E) of artificial intelligence (AI)-enabled and autonomous systems (referred to in this issuance as "DoD systems") acquired via the Defense Acquisition System or other non-standard acquisition systems.

DoDM 5000.101

Released Dec 9, 2024

DoD Instructions and Manuals

- DoDI 5000.98 (OT&E and LFT&E)
- Future 5000.DT
- OT&E/LFT&E Manuals (DoDM)
- DoDM 5000.96 (OT&E and LFT&E of Software)

High-level procedures on **WHAT** is required

Supporting Guides

Enterprise T&E Guidebook (August 2022)

- T&E Overview
- T&E for Major Capability Acquisition
- T&E for Middle Tier Acquisition
- T&E for Urgent Capabilities Acquisition
- T&E for Software Acquisition
- T&E for Defense Business Systems

https://www.test-evaluation.osd.mil/T-E-Enterprise-Guidebook/

Guidebooks and Companion Guides

- Cyber OT&E Guidebook
- Software
- Autonomous Systems
- **AI Systems**
- Modeling & Simulation

Includes detailed material for practitioners on **HOW** to implement aspects of the T&E Guidance

POLICY

GUIDANCE

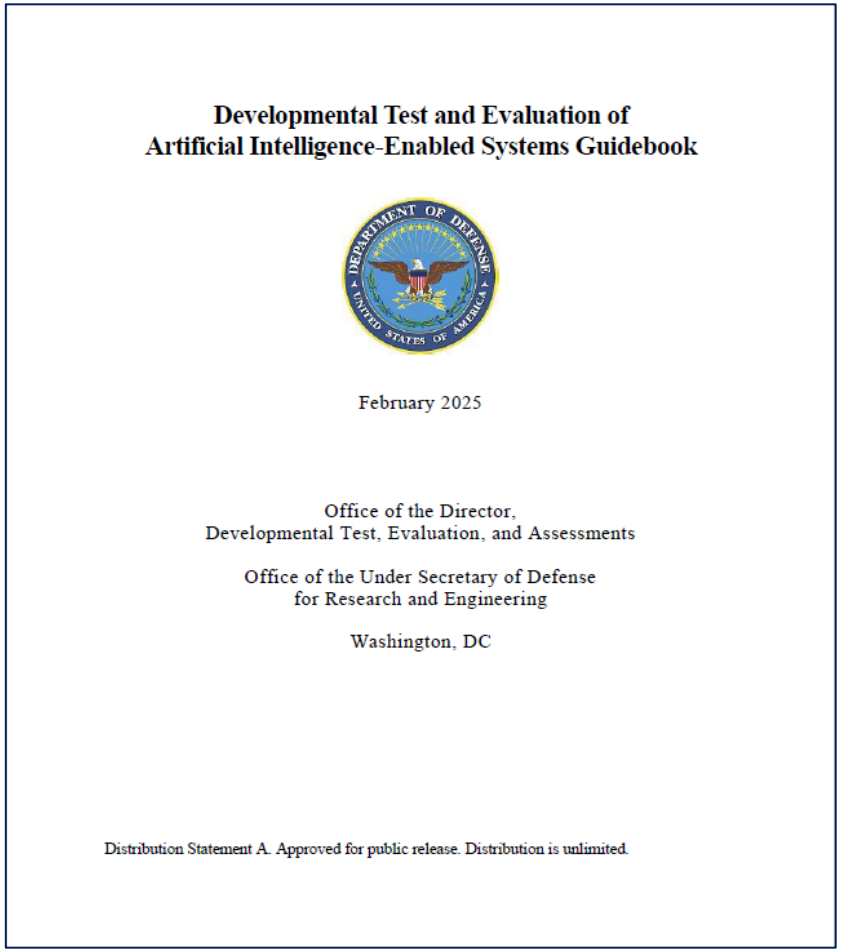
AI introduces challenges to traditional comprehensive testing approaches for components and systems

- System outputs cannot be comprehensively predicted over many iterations
- Small, sometimes even undetectable, changes in inputs alter outputs
- Many dimensional models produce outputs that are rarely repeatable
- AI links to training data makes the effect of corrupted data difficult to ferret out

Need for a T&E of AI Enabled Systems Guidebook!

Content of the Guidebook

Table of Contents



Developmental Test and Evaluation of Artificial Intelligence-Enabled Systems Guidebook

February 2025

Office of the Director,
Developmental Test, Evaluation, and Assessments

Office of the Under Secretary of Defense
for Research and Engineering

Washington, DC

Distribution Statement A. Approved for public release. Distribution is unlimited.

Chapter 2: DT&E of AI-Enabled Systems Overview

2.1 Introduction – AI-enabled Systems

2.2 DT&E Activities and Outputs

2.3 The CDAO T&E Strategy Frameworks

2.4 Implications of AI for DT&E Across the Life Cycle

2.5 Summary

“Comprehensive testing is no longer feasible for many AI components or AI enabled systems”

- Testing and measurement will remain paired to the models, but evaluation will expand beyond executed test conditions to characterization of capabilities, limitations, and risks of AI.
  - This will change cost and schedule of tests.
  - This will change the knowledge and actions of testers.

Chapter 3: AI-Driven Changes in T&E Practice

3.1 Introduction

3.2 ML Model Development and Assessment

3.3 Modeling and Simulation (M&S) for DT&E of ML

3.4 Use of Formal Methods in AI

3.5 Visibility into ML Models

3.6 Early Involvement

3.7 Summary

MODIFICATIONS NEEDED: Data Drives AI System Development Lifecycle

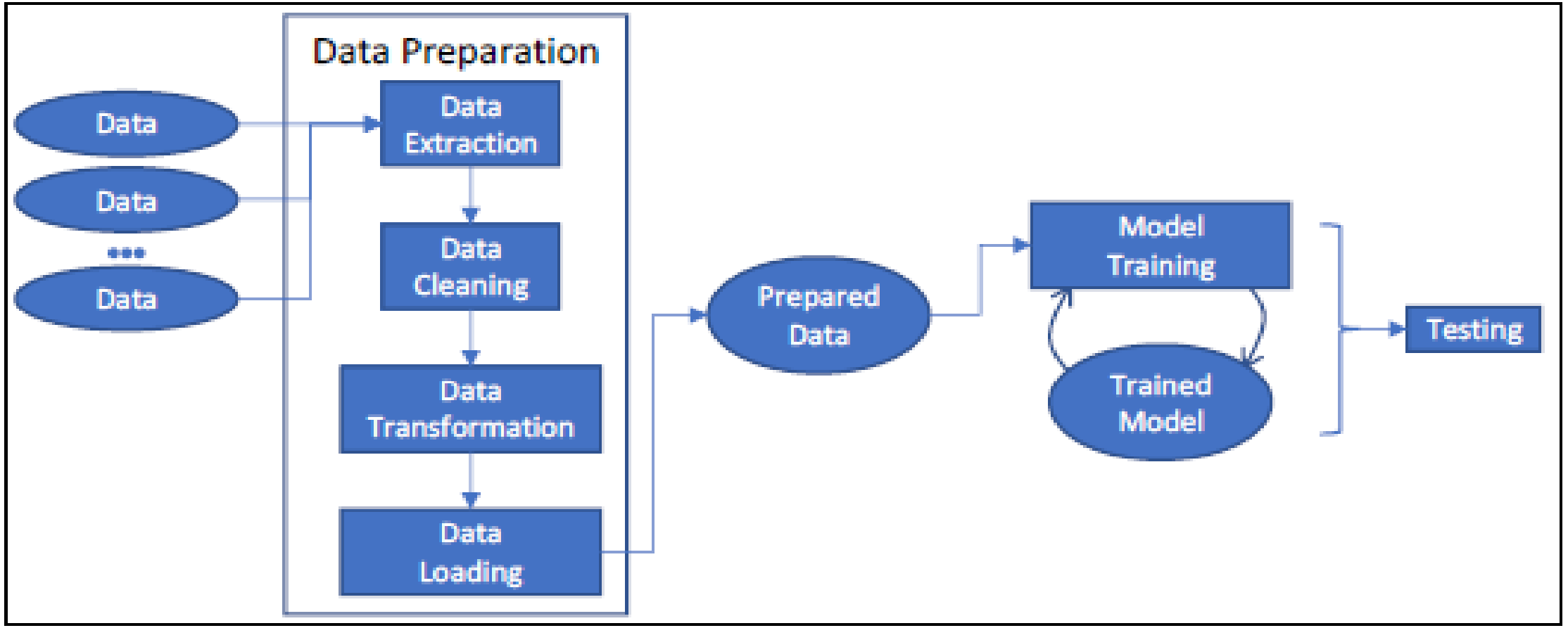


Figure 3-4. Notional AI System Development Life Cycle

Figure source: Developmental Test and Evaluation of AIES Guidebook, 2025, pg. 87

Verification, Validation, and Accreditation

Steps for the Verification, Validation, and Accreditation of AI map onto the AI Development Lifecycle

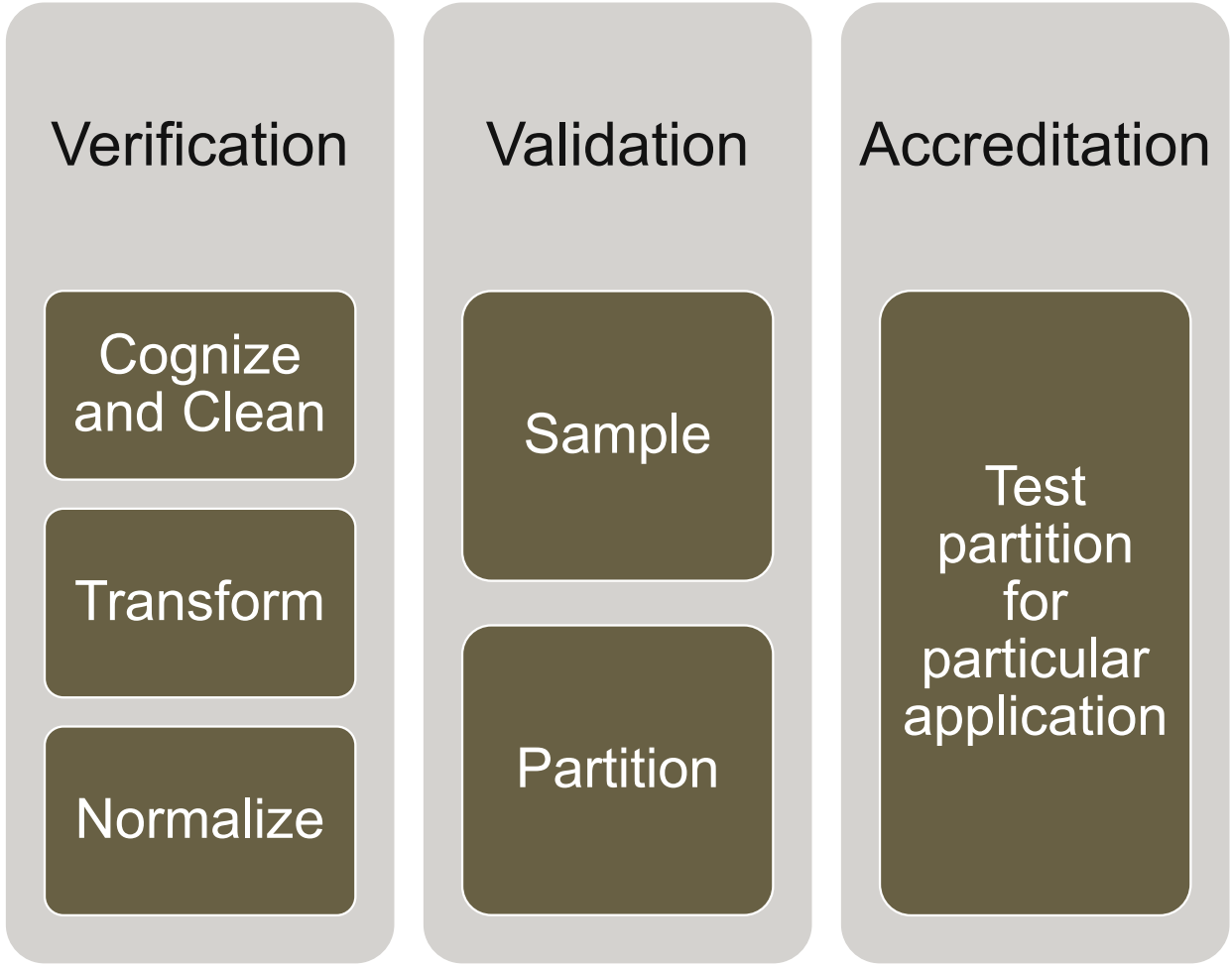


Figure source: Figure developed for this presentation

Augmenting with Synthetic Data requires additional considerations

Data augmentation with synthetic data invites additional concerns for T&E of AI. Synthetic data is generated via hand preparation or ML-model preparation.

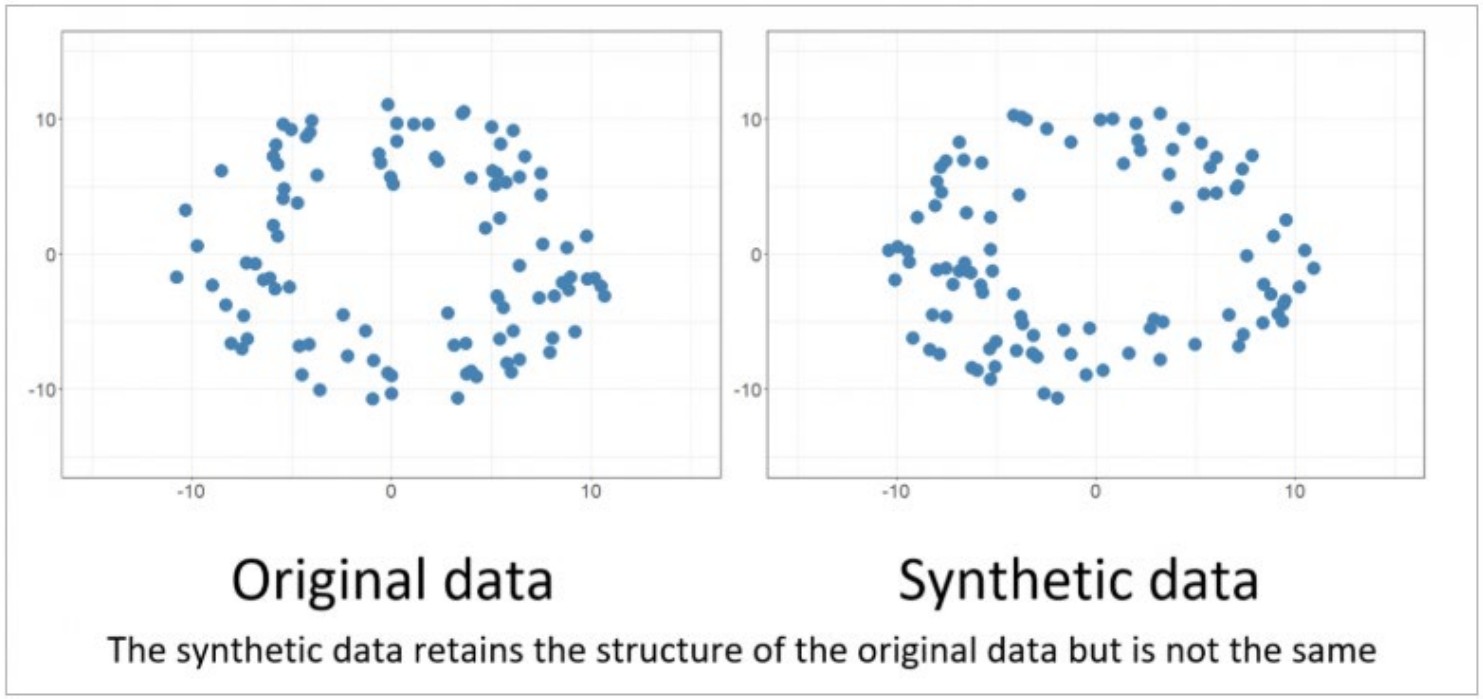


Figure source: https://dataingovernment.blog.gov.uk/2020/08/20/synthetic-data-unlocking-the-power-of-data-and-skills-for-machine-learning/

How does T&E of AI Incorporate Modeling & Simulation?

- M&S can substitute for open-air testing and also in test planning to optimize the information from the conducted tests.
- M&S on an ML model may not be practical but use of M&S data to support future ML development shows promise.

The Uses of Formal Methods

- Formal methods use mathematics and logic to rigorously evaluate how software will behave. It can be of particular use in early DT in proving that some classes of problems will not occur, obviating the need for some of the testing.

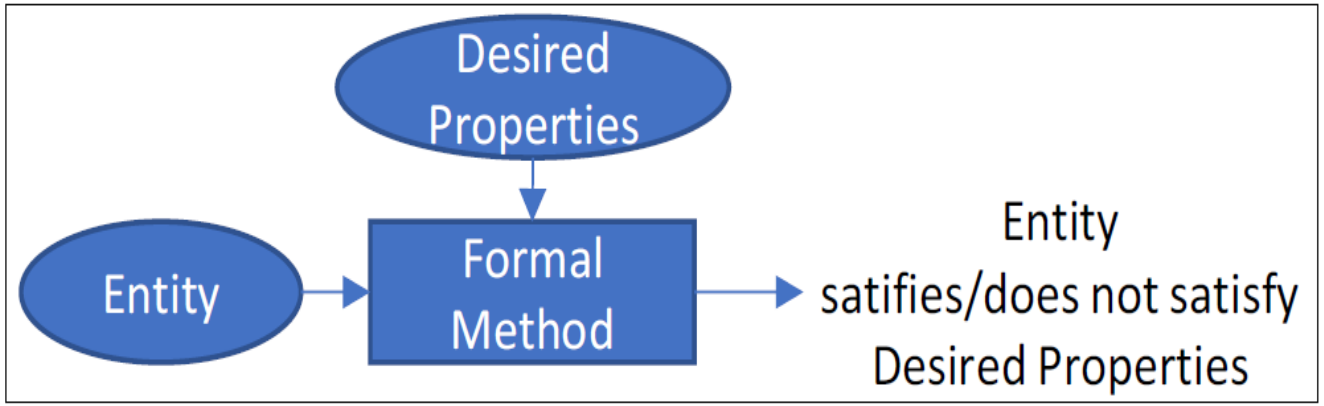


Figure 3-3. Formal Methods Concept

Defining Properties

Safety: Ensure the ML model does not produce unsafe outputs.

Robustness: Ensure the model is robust to adversarial inputs.

Fairness: Ensure the model does not exhibit biased behavior.

Correctness: Ensure the model produces correct outputs for given inputs.

Formal Verification

a. Model Checking: Use model checking techniques to verify that the ML model satisfies the specified properties. **Example Tools:** NuSMV, SPIN, PRISM.

b. Theorem Proving: Use theorem provers to formally verify the properties of the ML model. **Example Tools:** Coq, Isabelle, HOL.

c. Abstract Interpretation: Use abstract interpretation to analyze the behavior of the ML model at an abstract level. **Example Tools:** Astrée, Flawer.

Formal Validation

a. Testing: Use formal testing methods to validate the ML model against the specified properties. **Example Tools:** QuickCheck, Hypothesis.

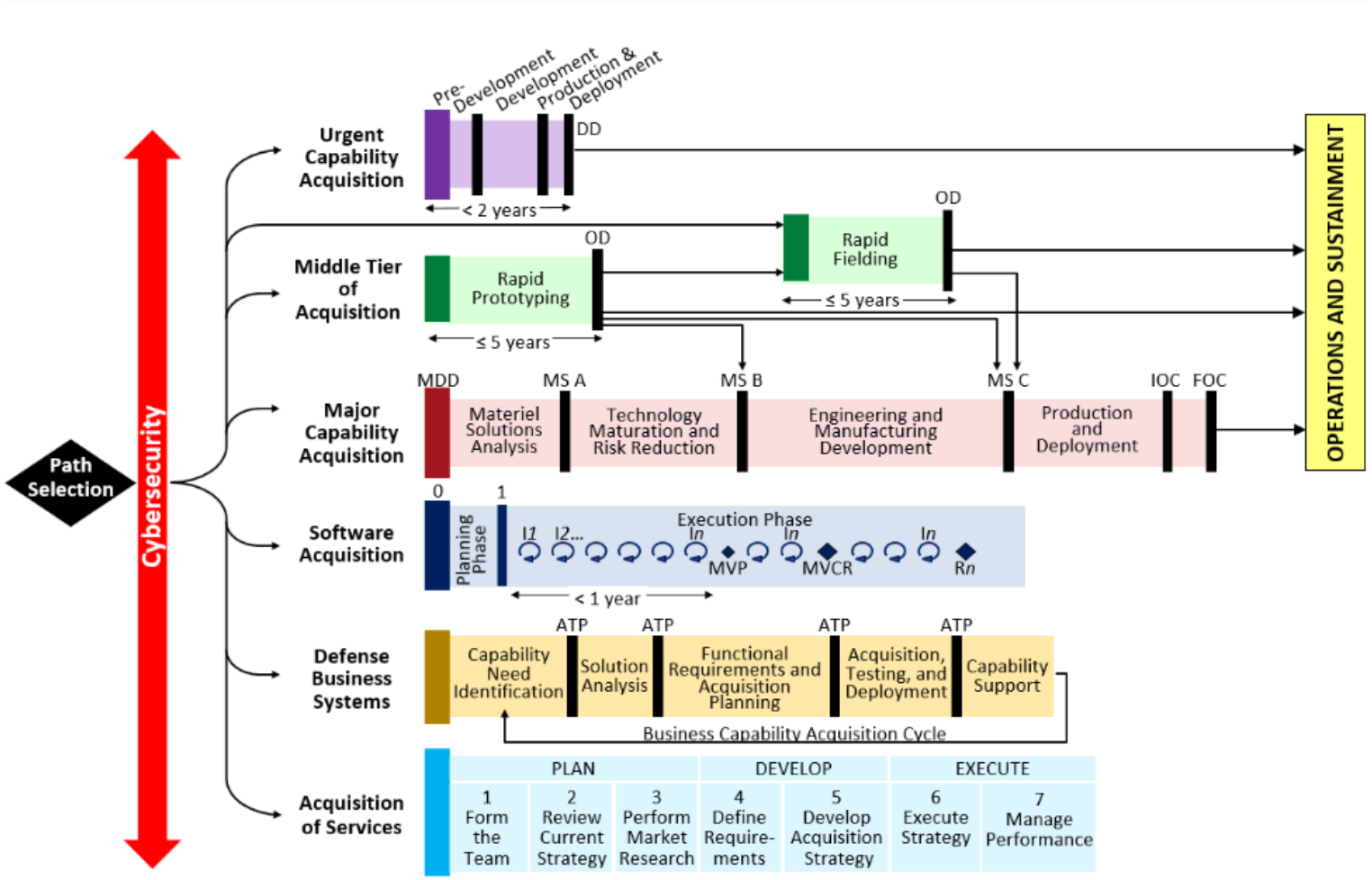
b. Runtime Monitoring: Use runtime monitoring to ensure the ML model behaves as expected during execution. **Example Tools:** RV-Monitor, Java-MOP.

Figure source: Developmental Test and Evaluation of AIES Guidebook, 2025, pg. 85

Visibility contributes to Explainability and Trustworthiness

- Visibility produces evidence that a system is sufficiently trustworthy for its intended use.
- Three components of a model's trustworthiness:
  - When employed correctly, the model will dependably do well what it is designed to do
  - When employed correctly, the model will dependably not do undesirable things
  - When paired with the humans it is intended to work with, the model will dependably be employed correctly

The Purpose of Early Engagement



Involving T&E teams early in AIES development enables mission-informed technology characterization.

A primary motivation for early involvement is to reduce the discovery of problems during DT, OT, or integrated activities late in development.

The focus should be on identifying which elements are likely to be stressed in operation. <u>The specifics will be system dependent.</u>

Figure source: DAU, Adaptive Acquisition Framework, https://aaf.dau.edu/

Chapter 4: Expanded Interactions for the T&E Community

4.1 Introduction

4.2 Contracting

4.3 Requirements Development

4.4 Concept of Employment (CONEMP)

4.5 Design Trade-offs

4.6 Accreditation and Certification Support

4.7 Summary

Chapter 4.2: Contracting – you get what you pay for

- Access to contractor data improves T&E
- VV&A of data and models is required by DoDI 5000.61
- Insight into development processes can contribute to assurance
- Instrumentation of the machine imposes a SWaP burden
  - Providing information of great value through development and deployment

- Contractor IP will have to be contracted for.

T&E professionals can advise about contracting for valuable information

Chapter 4.3: Requirements

- JCIDS lists four mandatory KPPs and two mandatory attributes
  - As defined, four are *unlikely* to be significantly influenced by the presence of AI, but still merit attention:
    - Force Protection KPP
    - System Survivability KPP
    - Energy KPP
    - Exportability Attribute
  - Focused attention is warranted for Interoperability and ‘sustainment’ broadly viewed
    - Interoperability Attribute – This would be affected if AI were guiding or mediating interface operation.
    - Sustainment KPP – the KPP addresses support to operational missions; ‘sustainment’ in the sense of keeping fielded AIES systems working will require monitoring and regression testing.

- KPPs, KSAs and APAs essential to the capability solution will generally need attention
  - Comparative requirements for AI performing human-like tasks are problematic
  - Emerging regulations can lead to de facto requirements, which are likely to appear in this category
    - E.g. confirming compliance with Modular Open Systems Architecture standards

Chapter 4.4: The CONEMP of an AI system must be baked into the lifecycle

"System design restricts possible CONEMPs for all systems. For **ML-enabled systems these constraints can become more complex as machines become less like tools and more like team-mates**. In addition, the iterative nature of design for ML-enabled systems can require iterations in the CONEMP to maintain or optimize performance."

Chapter 4.5: Design Tradeoffs

- Testing AI systems is challenging
  - For advanced ML systems the challenge can become a design driver
  - Procedural AI v. ML approaches
  - Use of run-time monitors to ensure behavior
- Providing Assurance to multiple stakeholders may also be challenging
- The impact of these challenges can depend strongly on the architectural and detailed design choices

Schedule impact or difficulty with approvals could drive design choices

Chapter 4.6: AI poses new Challenges for Accreditation and Certifications

- TVT data: What impact will data shortfalls have?
- HMI: What is need to establish governability and usability?
- ML brittleness: Can the sensitivity to small changes in input be adequately characterized to support accreditation/certification?
- ML robustness: Can the worst-case performance be adequately characterized to support accreditation/certification?
- Process: Is the development pipeline sufficiently characterized to support accreditation or certification?

Chapter 4.6: Cross walk: AI Challenges for Accreditation and Certifications

	RAI	Safety	Cybersecurity	Interoperability	DoDD 3000.09	Airworthiness	Data VV&A	Model VV&A
Quality of TVT Data	bias, privacy	coverage, representative-ness	data poisoning, data security		unintended engagements	coverage, representative-ness	bias, privacy, coverage, representative-ness, etc.	
HMT	governability	CONEMP		CONEMP, human factors	loss of control, unintended engagements	CONEMP		operator trust, explainability
Brittleness	reliability	predictability			unintended engagements	predictability		reproducibility
Robustness	reliability	worst-case behavior	worst-case behavior	emergent behavior	unintended engagements, loss of control	worst-case behavior	root causes of robustness failure	worst-case behavior
Vulnerability	governability	risk mitigation	adversarial inputs		loss of control	risk mitigation		
Model Development Pipeline	traceability		supply chain				synthetic data, normalization, etc.	objective function alignment

Figure adapted from Table 4.1, Developmental Test and Evaluation of AIES Guidebook, 2025, pg. 121

Wrapping up

- AI changes testers knowledge and test timing
- AI development increases opportunities for test engagement
- Language of VV&A and AI data can be aligned
- Synthetic data is an AI solution but can also be an AI problem
- Model visibility
- The concept of employment for AI is as a teammate

Get involved in what's coming Next!

Get Involved!: Next Steps for the DT&E of AI Enabled Systems Guidebook

- A 2nd edition is planned
- The 2nd edition will incorporate additional stakeholder feedback
  - We need the “view from the range” to ensure alignment to practitioner needs
  - We need more interaction with safety and sustainment user groups
- New and expanded content
  - Generative AI
  - T&E in risk management for AI systems
  - Reinforcement Learning
  - Formal Methods

Questions or Comments?

For questions or comments for this seminar, email: Sara Jordan (sjordan@ida.org) or Dave Sparrow (dsparrow@ida.org)

For engagement (on behalf of R&E) with future versions of this guidebook, email: (DTE_AIES_Guidebook@ida.org)