SLIDES ONLY

NO SCRIPT PROVIDED



# DoD Manual (DoDM) 5000.101

## T&E of AI-Enabled and Autonomous (AI&A) Systems

**Mr. Nilo Thomas**
**Director, Operational Test and Evaluation (DOT&E)**

***UNCLASSIFIED***

# DoD Instruction (DoDI) 5000.98, Section 3.3: OT&E

According to the overarching DoDI 5000.98, Operational Test Agencies (OTAs) will conduct OT&E of artificial intelligence enabled DoD systems in accordance with DoDM 5000.101.

**Goal**: Improve test planning, test rigor, and implementation of leading practices for identifying and quantifying risks of AI-enabled systems, and characterizing system performance across the acquisition lifecycle.

**Context:** Beginning to test oversight systems with AI/ML and experimental autonomous capabilities. It will take time to reach authoritative best practices, T&E activities, and guidance.

## PREVIOUS POLICY AND GUIDANCE

**DOT&E is developing guidance to convey new expectations to PMs and the Services.**

Unlike other DoDMs, DOT&E has not previously provided policy or guidance on AI. The DoDM includes directed and implied test activities; the guidebook will make expected test activities more explicit, with timelines for application.

# DoDI 5000.101
# AI&A
## Overview

# What is New, Updated, or Re-Emphasized in DoDM 5000.101?

This slide highlights the major shifts within the DoDM, relating to planning, content or documentation, and deliverables.

## PLANNING

- Science and Technology-Based AI testing
- Supports V&V of AI datasets and models
- T&E of AI model behavior
- Human-system integration and human machine teaming
- Adversarial and counter AI testing
- Five DoD AI Ethical Principles
- Resourcing considerations for AI

## CONTENT / DOCUMENTATION

- Datasets and metadata used to train and test AI models
- Sustainability and data pipeline update plans
- Tracking of system safety and unexpected behavior
- Ensuring access to system safety assurance, software and hardware assurance, and mission performance assurance documentation

## DELIVERABLES

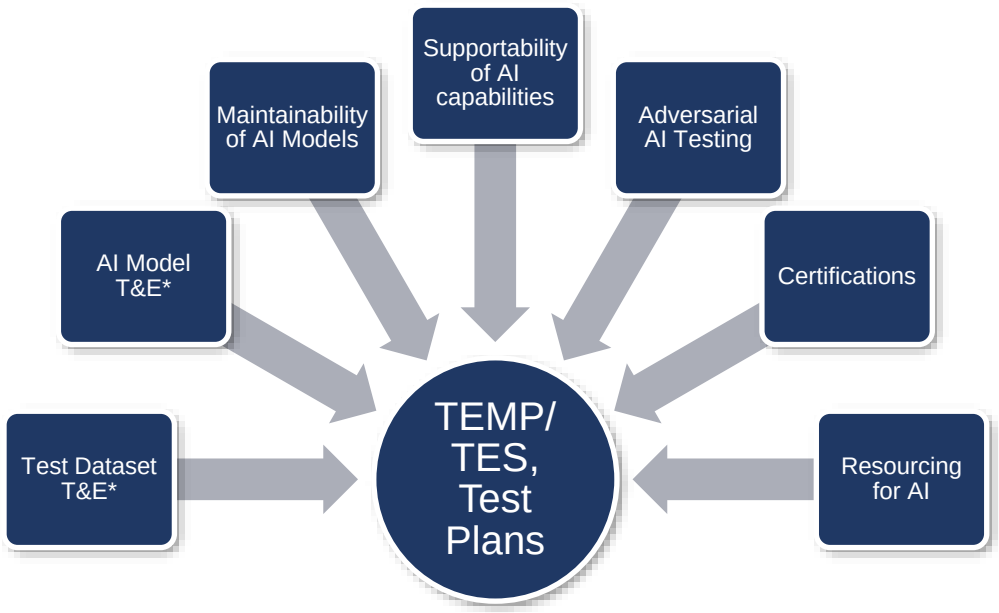
- Data management and governance plans
- Data cards
- Model cards

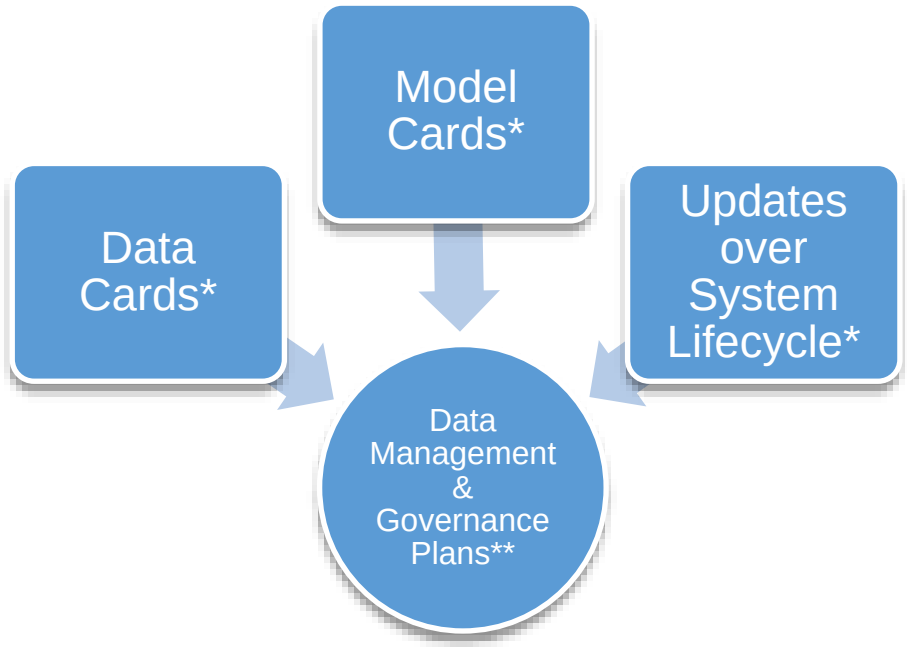
NEW

REINTERPRETED FOR AI

# What Will the T&E Community Need to Do Differently?

The introduction of AI capabilities influences the scope, scale, and types of testing needed.




*Test activities may be executed as part of contractor, developmental, and/or integrated testing. Relevant information should be documented and shared for operational test planning, execution, and reporting.

**Data management plans are required by DoDI 5000.98.

# Risks this DoDM Addresses

Changes are not necessarily one-size-fits-all programs, and DOT&E understands that guidance will need to be tailored. We plan to work at the program-level to understand what works, what does not work, and to revise guidance appropriately.

## Concerns and Risks

- Based on how programs currently write about AI&A capabilities in TEMPs, TESs, and test plans, it can be unclear if and how **novel AI&A risks** are being identified, characterized, and accounted for
- **Datasets and ML algorithms** used to build AI&A capabilities have direct impacts on system performance, risks, and test needs
- Leading practices for **HSI**, cyber survivability and attack surfaces (e.g., **adversarial and counter AI**), and **responsible AI** methods for AI&A are rapidly changing

## Expectations with a Signed AI&A DoDM

- As program test documents are developed and revised, programs should provide additional AI&A details
  - AI and autonomous capabilities should be adequately described in **system descriptions, CONOPS,** and include **new operational users and maintainers** (e.g., data scientists, ML engineers)
  - Shared documentation on datasets ("**data cards**") and AI models ("**model cards**") will assist planning, scoping, and executing AI&A lifecycle testing
  - DOT&E will work with T&E stakeholders to disseminate **leading practices** for T&E of AI&A

# Risks this DoDM Addresses

## AI&A System Test and Risk Considerations

- **Training datasets** impact performance and should be validated for their system's use cases; government testers have equities into a **test dataset's*** splits and use
- **AI model testing** processes need to adapt based on frequency of retraining and other updates
- Users/operators must be able to work with more complex, potentially non-transparent systems, impacting **HSI considerations**
- Ensure continued, **robust AI-enabled performance** with updates and new missions
- DEPSECDEF declared that all DoD AI must meet **5 AI Ethical Principles**

## New Responsibilities and Topics of Interest

- WIPT/ITT should account for new AI&A risks and test needs
- **Data management plans** account for AI considerations
- Risk management needs to account for **changing AI/ML models** in systems

* "Test dataset" refers to data put through AI models to assess and characterize model performance; it's distinct from test data collected during test activities

# DoDM 5000.101
Contents

# Evaluation of Datasets: Verification and Validation for Use

## Testing (and Training) Datasets

- AI/ML testing datasets are **distinct from "test data"** collected at test events
- **AI/ML testing datasets are data pushed through models to assess performance**
- Datasets should have:
  - **Representativeness**: datasets should represent data the system is expected to see in field
  - **Independence**: the government's testing datasets should be independent of vendor training data
  - **Coverage**: datasets should cover the full operational space for operational relevance
    - Multiple metrics are available
    - Prioritize areas of less coverage for testing

## New Responsibilities and Topics of Interest

- Created by AI/ML developers and updated over the system's lifecycle
- Data cards detail provenance, coverage, and other relevant details
- Datasets should update over time; details of data cards should update with them
- **Risk**: data used for training bring new supply chain and safety vulnerabilities that are unaddressed
- **Risk**: data used for training is not operationally relevant, look "good enough" during OT, and prove ineffective and/or fail upon fielding

# Evaluation of AI/ML Models

## AI/ML Models

- AI/ML models will go through test, evaluation, verification, and validation *at the component-level (likely CT, DT)*
- Ensure relevance of component tests to operational testing is known
- Characterize model performance, biases, robustness, scalability, and security, among other aspects[1]
- Increased emphasis for **mission-based metrics** that can be tied to planned OT
- **Risk**: model performance and effectiveness issues pushed to live or other, more expensive testing
- **Risk**: testing does not capture edge cases, failure modes; first see in fielded systems

## Model Cards (New Documentation)

- Should be developed early and updated over time
- Detail provenance (e.g., adapted from public model), testing, and performance
- Detail planned and actual model updates over time
- CDAO and TRMC have prototypes that they are iterating into standards with other stakeholders
- **Risk**: AI/ML models bring new supply chain and safety vulnerabilities that are unaddressed

[1] Details of testing AI models are beyond the scope of this presentation

**Acronyms:** CDAO: Chief Digital and AI Office, TRMC: Test Resources Management Center

# Evaluation of Human Interaction With AI&A Systems

Various stakeholders have differing evaluation needs for AI&A system performance: Testers, commanders, programs, and operators/end users may have different needs.

- **Interpretability**: users need some level of understanding how AI&A functions to trust it
- **Explainability**: users may need clear and valid explanations of AI/ML decisions made
- Assess whether user's **trust is appropriately calibrated** across the operational envelope (not over- or under-trusting)
- Test events may need demonstrations of identifying problematic performance and executing mitigations, e.g., model changes
- AI systems need to be **governed** adequately to retain responsible positive control: demonstrate users detecting when and how to roll back systems to prior, more stable versions (under problematic performance)
- **Risk**: operators do not have information to govern systems and prevent unintended engagements
- **Risk**: commanders do not have enough assurances to allow AI&A on ranges and employ in operations

# Evaluation of Learning Systems That Change Over Time

## AI/ML Models Inherently Change

- When frequent AI updates are planned, OT may evaluate the adequacy of **ML model maintainability and supportability**
- Verify processes are in place to detect, track, and respond to deviations in performance
- Assess adequacy of retesting plans upon AI/ML model redeployment

## DoDD 3000.09[1]

- DOT&E is responsible for:
  - **Establishing standards** for data collection post-fielding for monitoring and assessment by programs
  - Coordinating with PMs and appropriate military commanders to identify when additional T&E is required to prevent unintended engagements or resist adversary interference
    - Primarily will occur due to system design or operational environment changes (e.g., System to operate in CENTCOM and will be employed in INDOPACOM)

[1] Department of Defense. “Autonomy in Weapon Systems.” DoD Directive (DoDD) 3000.09. Washington, DC: OUSDP, 25 January 2023.

# Other Practices in the DoDM

## Multiple New Responsibilities and Topics of Interest

- WIPT/ITT has new responsibilities relating to AI&A risks and test needs
- Data management plans account for AI considerations
- Risk management needs to account for changing AI/ML models in systems

## CDAO AI T&E Frameworks Include Additional Guidance On Implementation

**DTE&A**

- AI model testing
- System integration testing
- Human-systems integration testing
- Operational testing

**DOT&E**

**Responsibilities**

# Reflections – What Does This Mean for DOT&E AO Program Oversight?

Reviewing TEMPs/Test Plans once this guidance is issued

- Identifying User and operator HIS, HMT, and trust considerations
- Ensure continued, robust AI-enabled performance with updates and new missions
- Ensuring DoD AI systems are meeting the 5 AI Ethical Principles
- AI models change over time and the test process must adapt to the changes in the model over time
- Data Management Plans to include AI&A considerations
    - Understanding Test and Training datasets and how AI&A programs use them
    - Having access to the AI model and data cards to determine adequate AI scoping
- Aware of advanced threats to AI&A systems due to adversarial AI techniques

CLOSING STATEMENTS

# Q/A

THANK YOU