# DTE&A Systems Engineering Processes to Test AI Right

## OSD DTE&A MITRE Support Team

### September 2023

Sponsor: OUSD DTE&A
Project No.:101074.23.401.D320.P04

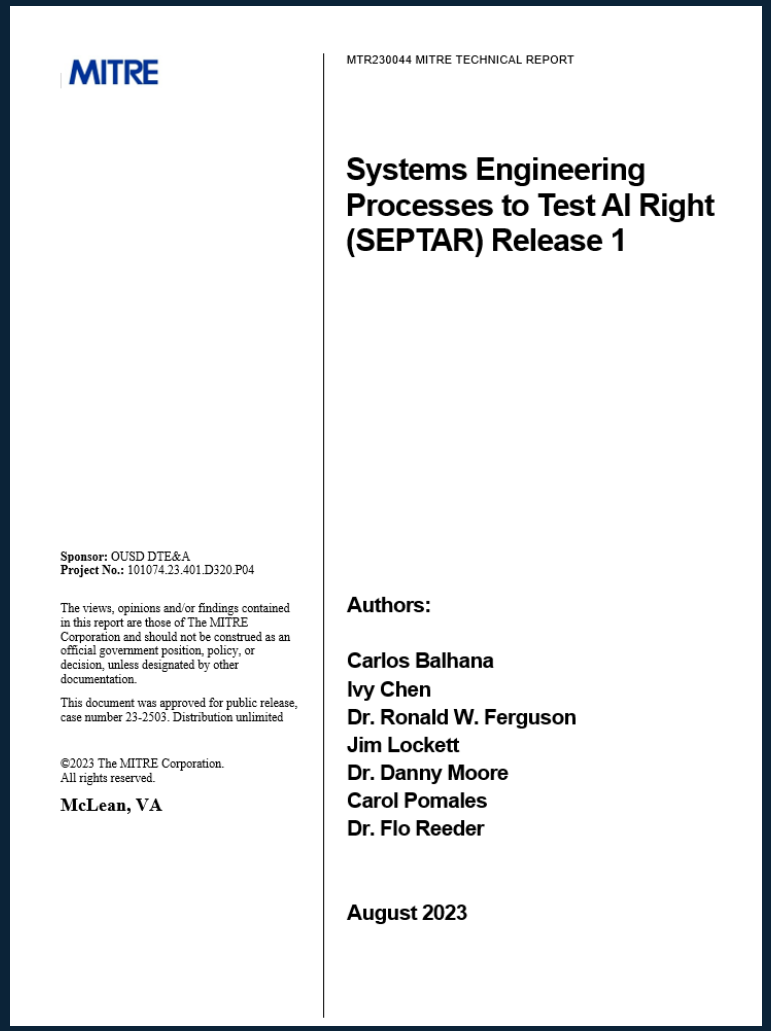


McLean, VA

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD

# SEPTAR: Objective

- The DoD is making sizable investments ($14.7 billion) in Artificial Intelligence (AI) R&D and acquiring AI through programs
- SEPTAR was developed by a MITRE team supporting Developmental Test Evaluation and Analysis (DTE&A) with input from DOD Stakeholders and SMEs across DOD (DOT&E, CDAO, DTE&A, TRMC, ATEC), academic partners (V. Tech, GTRI, CMU, JHU/APL) and FFRDCs (MITRE, Aerotek, IDA, STAT COE).
- It presents the benefits for proactive planning for Test and Evaluation (T&E) activities for AIES
- SEPTAR recommendations were determined to deliver the following goals:
  - Ensure AIES are more likely to be delivered on time
  - Meet budgetary goals,
  - Perform effectively to meet mission expectations
- Assuring and understanding the processes used to build the AIES informs on the later T&E


MITRE

MTR230044 MITRE TECHNICAL REPORT

**Systems Engineering Processes to Test AI Right (SEPTAR) Release 1**

**Sponsor:** OUSD DTE&A
**Project No.:** 101074.23.401.D320.P04

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

This document was approved for public release, case number 23-2503. Distribution unlimited



**McLean, VA**

**Authors:**

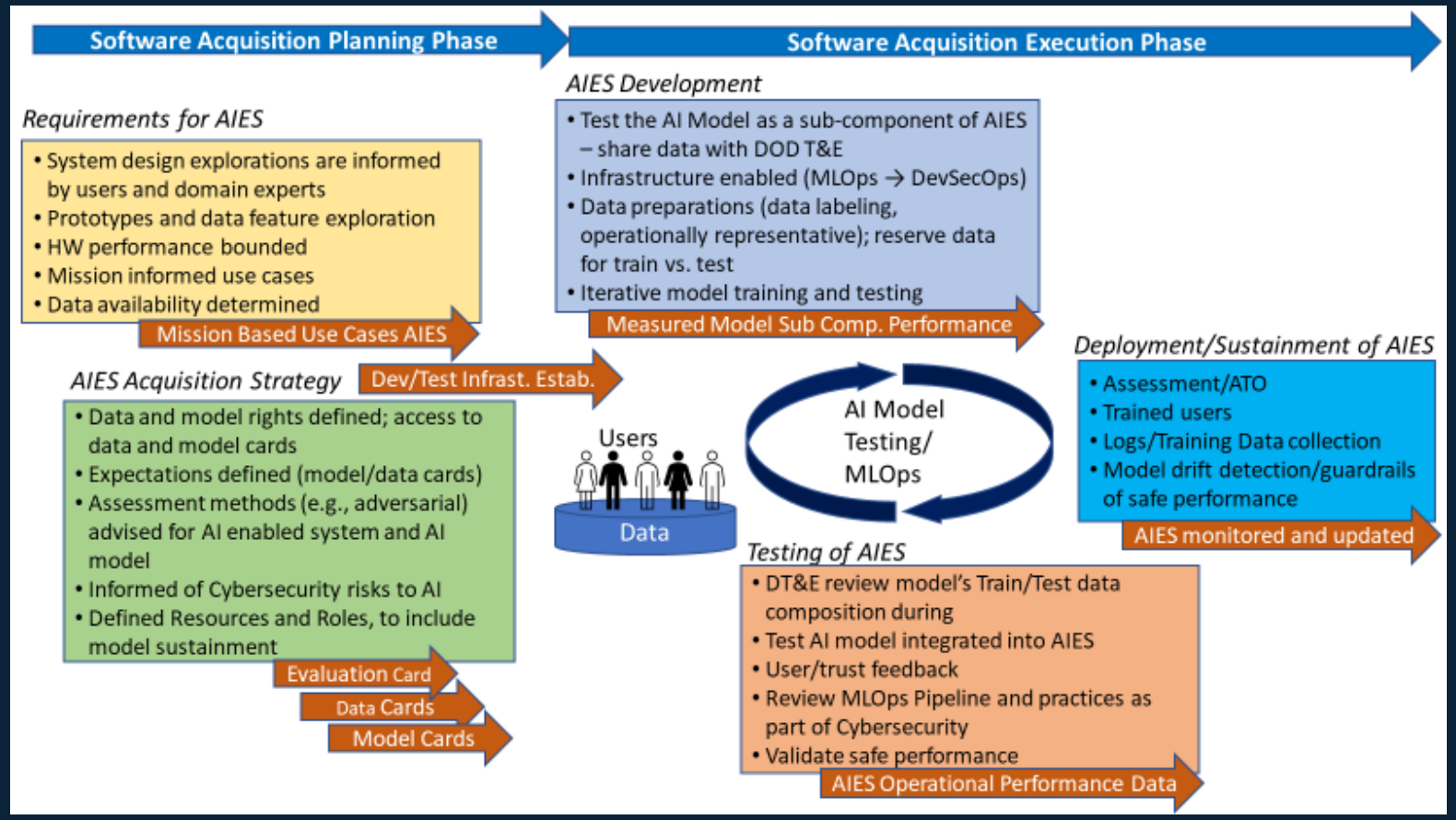
**Carlos Balhana**
**Ivy Chen**
**Dr. Ronald W. Ferguson**
**Jim Lockett**
**Dr. Danny Moore**
**Carol Pomales**
**Dr. Flo Reeder**

**August 2023**

*Note: For all references and citations from these slides, please see the SEPTAR Paper*

# Pilot Planning and Considerations

- DTE&A is looking for 4-5 programs to partner with  to share emerging guidance in testing of T&E of an AI-Enabled system
- Work with program management & T&E staff to advise of recommendations and provide follow up execution support
- No additional charge to the program and lessons learned from the experience will allow DTE&A to shape future guidance and T&E capability enablement



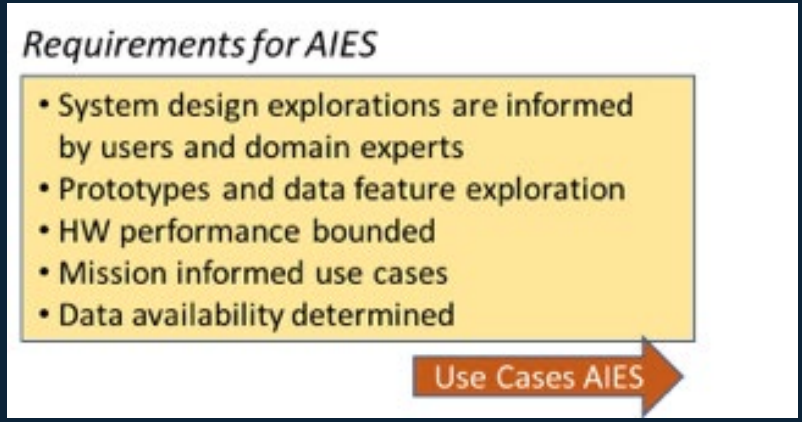

**Goals**

- Broadening the T&E continuum
- Defining data needs for AIES up front
- Evaluating the SELC to inform AIES trustworthiness

MITRE

# Requirements Summary

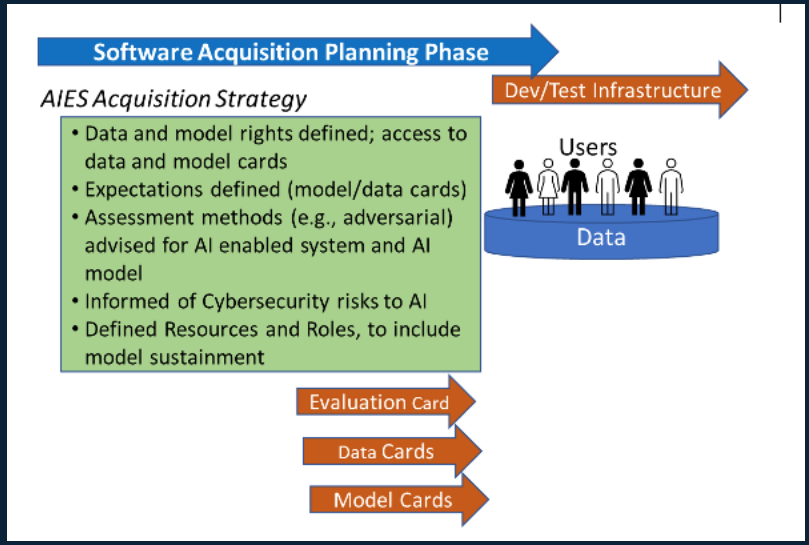
- Create Agile teams early in process
- Involve Testing SMEs before procurement
- Develop mission informed use cases
- Create and validate mission informed use cases considering user needs and stakeholder impact
- Utilize early prototyping/proof-of-concepts
- Account for Human Machine Teaming methods
- Document requirements for T&E
- Address unique AI infrastructure and hardware requirements early on

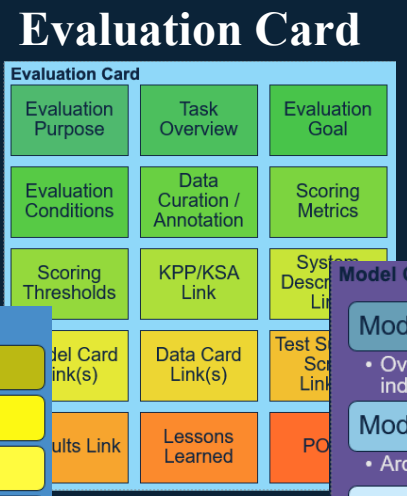


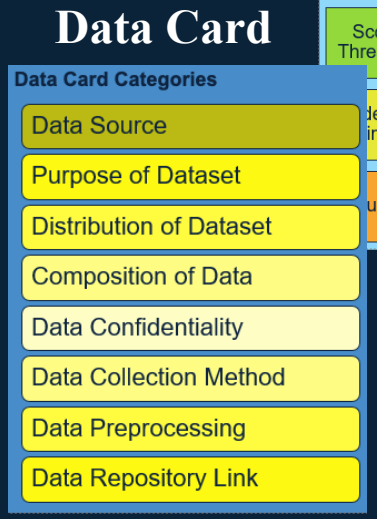

- ✓ Enables quick iteration and encourages adoption
- ✓ Ensures selected AIES meet necessary requirements
- ✓ Provides clear context for T&E
- ✓ Considers user needs and stakeholder impact
- ✓ Identifies potential challenges and limitations
- ✓ Ensures effective implementation
- ✓ Establishes clear evaluation criteria and objectives
- ✓ Ensures accurate evaluation

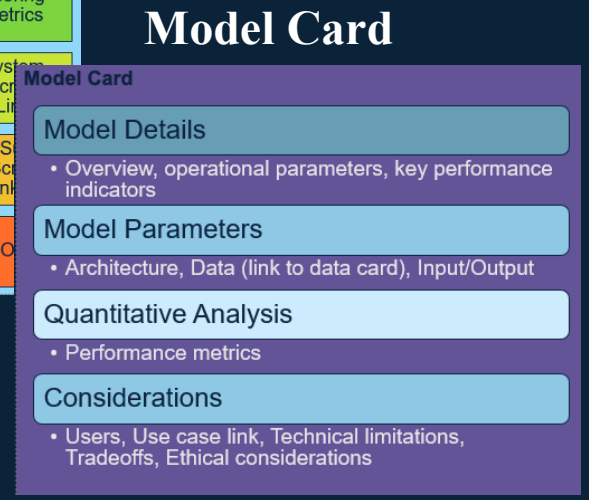
# Acquisition Strategy/Contracting Summary

- Implement key agreements in acquisition strategy
- Ensure delivery of documents CDRLs (Contract Data Requirements Lists) to DOD
- Enable T&E processes with Data Cards, Model Cards, and Evaluation Cards
- Address data quality and data rights
- Be informed on threats to AIES



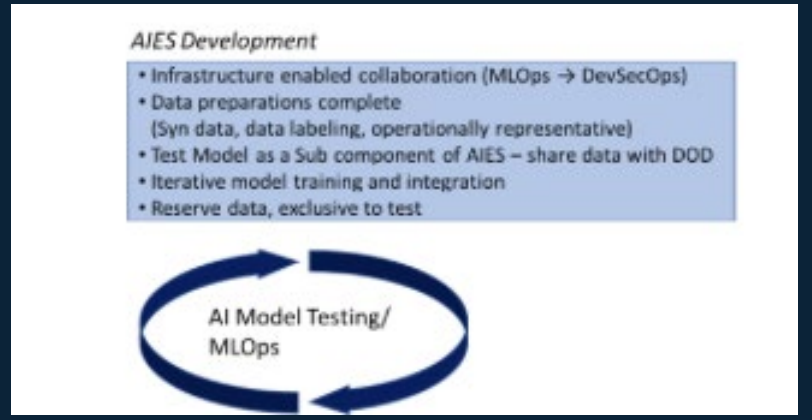



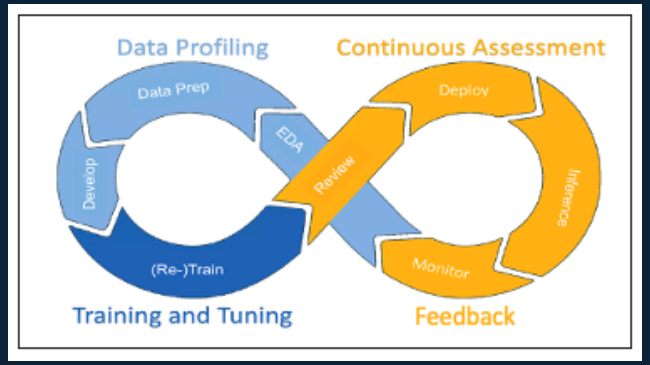

- ✓ Enables a more effective evaluation of AIES
- ✓ Facilitates more comprehensive assessments
- ✓ Conducts T&E using accurate and reliable data
- ✓ Identifies potential vulnerabilities and risks earlier
- ✓ Enables proactive mitigation strategies
- ✓ Ensures the robustness of AIES

MITRE

# Development & Model Testing Summary

- Treat the AI Model as a sub system with its own unique testing requirements that are reported to Test Stakeholders
- Iterative process that requires multiple cycles to define the right AI model for the provided data
- Development Phase: AI Model and AIES development, testing, and training
- Leverage and update Evaluation Card, Model, and Data Cards
- Enable T&E community to understand AIES
- Contribute to total body test evidence
- Ensure proper data protection and reserve key data

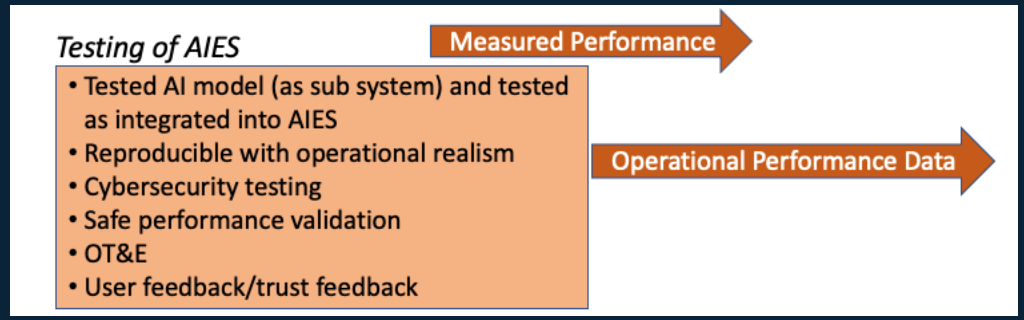



*The Machine Learning Operations (MLOps) framework provides a system-oriented approach specifically tailored to AIES, that may facilitate more effective T&E of the AI Model*

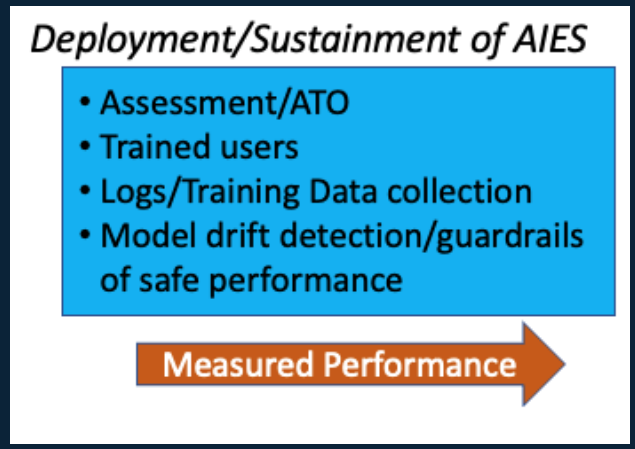
MITRE

# Test and Evaluation Summary



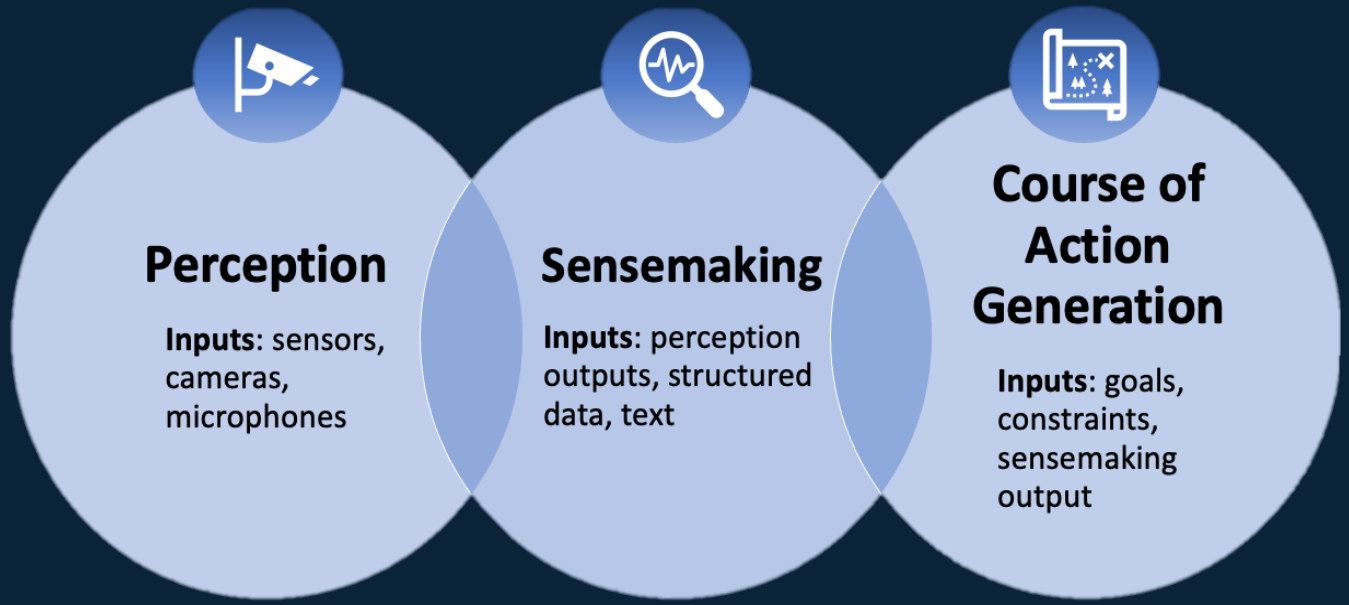

- Use workshop process for T&E planning phases
- Advocate CDAO processes for key consideration including those in CDAO T&E Workshop and next release in September 2023
- Conduct formal DT&E and OT&E after model T&E in development phase
- Leverage mission-informed use cases in T&E phases
- Address AIES-specific threats with cybersecurity testing (VOLT/Validated On-line Life-cycle Threat) threat assessment
- Consider level of access and understanding of models (e.g., black-box systems)
- Discuss specific AIES performance measures (e.g., HMT)

# Deployment/Sustainment Summary




- Maintain evaluation cards during Deployment/Sustainment
- Continue capturing specifications, processes, and results
- Integrate reporting components with CD (Continuous Delivery) test tools
- Evaluate and optimize AI model and pipeline iteratively
- Address emerging mission needs and evolving fielding conditions
- Establish recurring user feedback elicitation capabilities and activities
- Identify user support gaps and barriers in CD life cycle
- Develop a digital reporting framework for issue tracking and resolution

# SEPTAR Use Cases




- **One use case from each category – describes T&E impact throughout SE lifecycle for that type**
- ***Perception*: Tank recognition in images**
- ***Sensemaking*: Document triage of Arabic hardcopy**
- ***Course of Action Generation*: Battlefield course of action generation training**

***Perception*** *covers a range of recognition or categorization tasks, such as automated speech recognition, computer vision for image recognition, and identification of objects/elements from sensor data. Since perception tasks, by and large, are categorization tasks; they have the ability for ground-truth evaluation, which is based on knowing the output the system should produce for a given input. Where they tend to have a greater T&E design burden is in the variation of the sensor or information capture tools, such as microphones or cameras. The use case is one of image recognition or identifying objects in images.*

***Sensemaking*** *tasks include a wide range of data transformation, classification, and combination activities. Information extraction, machine translation (MT), social graph analysis, and event prediction are all types of applications in this category. Typically, these applications rely on ground truth (or near ground truth) reference values with which to evaluate model performance.*

***Course of action generation*** *is in the class of applications where the AI-based system will consider a wide range of options in combination to achieve a goal. The options typically have constraints that help determine their viability as selections. A classic example is logistics planning where systems help determine the optimal route selections and load balancing to enable package delivery services to meet deadlines and work efficiently. Recently, COA generation has expanded to include identifying patterns-of-life from multiple non-traditional sources or to support modeling battlefield possibilities. The use case selected in this category is COA recommendation for military mission planning.*

MITRE

# Pilot Benefits to PMOs

## Pilot Window: May – September 2023+

**SEPTAR:**

- Requirements
- Acquisition
- AIES Development
- T&E of AIES
- Deployment
- Sustainment

PMO's planning will be informed on considerations that need to be addressed for this new technology, answering their calls for SME

Insights gained from prototypes is used to defining use cases (including negative ones) improving performance, cyber resiliency and avoiding undesirable performance drop offs

Informs contract language to ensure CDRLs are accounted for and delivered preventing schedule or cost overruns to enable rapid T&E

When AI Training Data is well designed (e.g., operationally relevant) and employed successfully it can prevent bias and ensure AIES operational effectiveness

Metrics to measure AI performance during DT and OT are understood and employed to facilitate independent T&E

HSI is included in design and measured to ensure user adoption and trust

Deployed AIES are monitored (data collection in place) for drift and retraining is applied using standing pipeline/infrastructure

Users are trained and apply appropriate trust to the AIES

- ✓ AI is an emerging practice and uninformed practices can lead to unsuccessful AI implementations
- ✓ PMOs can leverage recommendations built from SMEs (MITRE, STAT COE, IDA, CMU, V Tech, GTRI, JHU APL) and SME from OSD, (DTE&A, DOT&E, CDAO, A&S)
- ✓ Positive PMO engagement demonstrates to A&S the intent to apply thoughtful AIES practices
- ✓ Helping to inform future policy and guidance and Projects on DTE&A oversight can avert later challenges

MITRE

# Next Steps

- *Define additional Government lead(s) for coordination and project(s) to participate in pilot*
- *Define focus area for pilot*
  - Requirements/Acquisition
  - AIES Development/T&E of AIES
  - Deployment/Sustainment
- *Setup workshop to deep dive options and finalize approach*
- *Establish timeline and periodic check points*
- *Survey success/lessons learned from application*
- *Update SEPTAR to inform DTE&A T&E of AI Policy and Guidance*

# Backups

# The Cards

- Model Cards define the model: description of the model and usage,  owner/license, measures of quality/performance, data used to train the model, intended users, limitations (including ethical)
- Data Cards come from the data owner/provider.  It defines the data: source, collection, licenses, sanitization, labeling
- An AI model's evaluation card documents the methods and conditions under which the AI models will be evaluated and tracks results of these AI-subcomponent level assessments. The evaluation card approach ensures a consistent and well understood T&E methodology is used.
- These cards should begin during the earlier planning phases (requirements, acquisition strategy) and be updated frequently throughout the SDLC.
- Data and model cards must be reviewed and verified by the T&E team throughout the SDLC process and any associated evaluation cards must be updated to reflect any changes made to the data and model cards over time.