CDAO

Chief Digital & Artificial Intelligence Office

Test and Evaluation of AI Enabled Capabilities

Distribution Statement A
Approved for Public Release; Distribution Unlimited

Disclaimer: The appearance of company names or vendors within this presentation does not imply or constitute DOD endorsement.

CDAO Assessment and Assurance Mission

Provide stakeholders with justified confidence that DoD AI-enabled systems meet requirements and support mission through ethical action.

Stakeholders include warfighters, commanders, program managers, acquisitions, regulators, taxpayers, international allies

1. **Assurance Best Practices**
2. **Assurance Capabilities Development**
3. **Program Assessment**

CDAO

Assurance Best Practices

CDAO

Focuses on working-level testers

Empower testers without AI/ML expertise to reach an ~80-90% solution

T&E Strategy
Educate testers about concepts and AI-specific concerns

Test Plans
Explain existing techniques, develop new ones

Negotiating
Empower testers by implementing best practices as policy

Products
Provide repos, widgets, and guides for common tasks

↓

JATIC

CDAO

CDAO T&E Frameworks Overview

Operational T&E (OT&E)
Evaluating an AI enabled-capability (AIEC) performing representative missions in a realistic environment against realistic adversaries

Human Systems Integration (HSI) T&E
Evaluating an AIEC’s ability to help stakeholders observe and orient to their environment, make informed decisions, and carry out their missions.

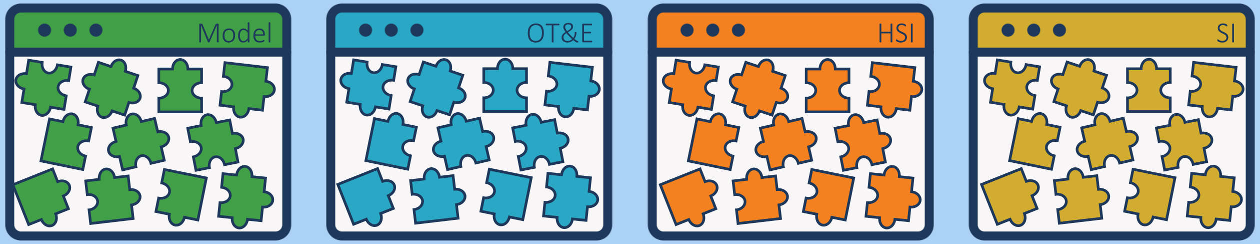
Systems Integration (SI) T&E
Evaluating the reliability, functionality, interoperability, compatibility, and security of an AI model within a system.

AI Model T&E
Evaluating and documenting AI models and data across performance dimensions informed by system and mission constraints.

CDAO

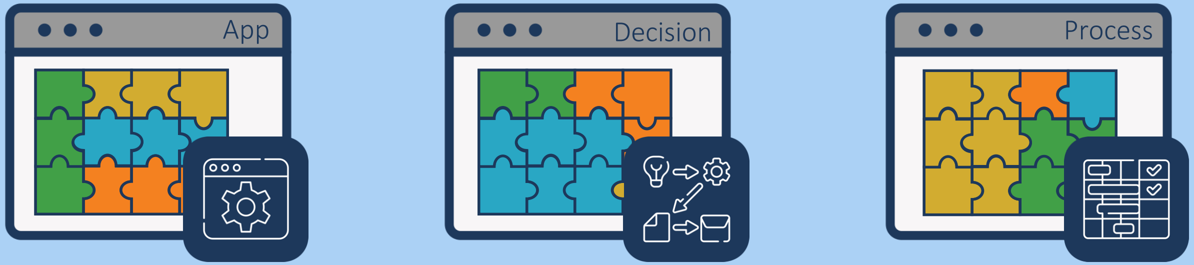
The Framework Vision

Test & Evaluation Strategy Frameworks



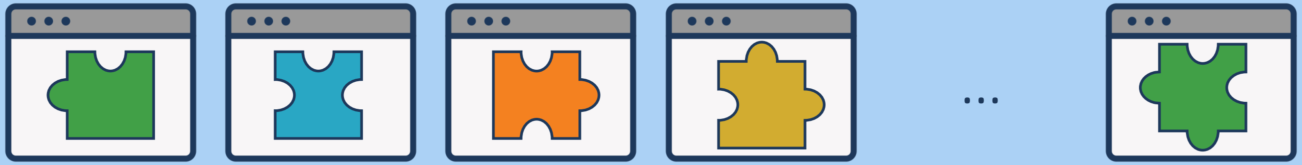
Foundational understanding for AI and/or DoD T&E novices

Use Case Guidebooks + Codebooks



Concrete, tailored guidance mapped a particular use case

Concept Deep Dives



Technical details of T&E methods and metrics mapped to tradeoffs

CDAO

AIEC characteristics can exacerbate preexisting challenges

Complex Decision Making

Black Box Algorithms

Gamification & Reward Hacking

Agile, Iterative Development

Overfit to Training Data

“Shift Left”

An ounce of prevention is worth a pound of cure.

“Shift Right”

T&E cannot stop at deployment.

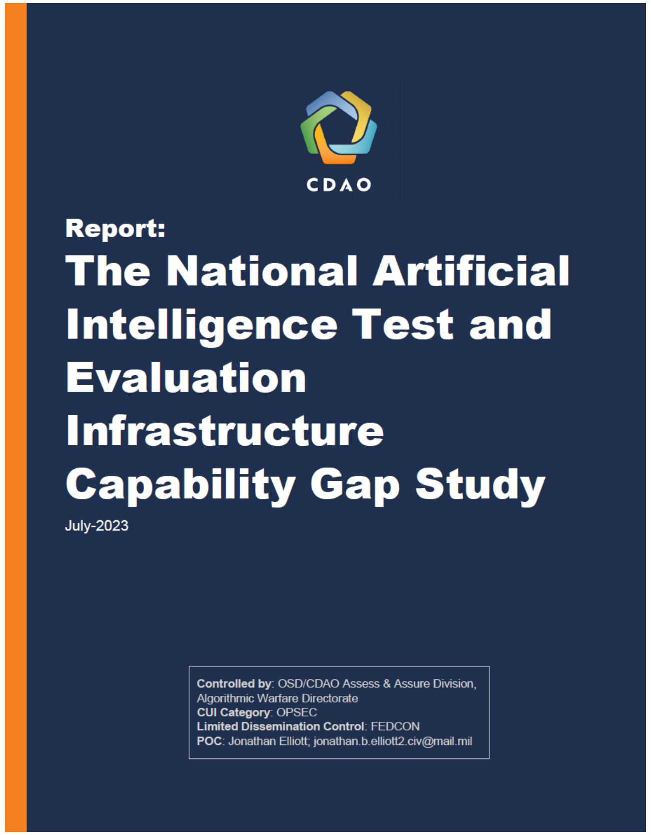
Assurance Capabilities Development

CDAO

Report on AI T&E demand and gaps

- “There is widespread interest for DoD enterprise-level T&E infrastructure to address the novel and exacerbated challenges posed by the T&E of [AI].”

- “While programs are currently investing locally in T&E resources… there is still a consistent desire across survey programs for DoD enterprise support.”



CDAO

Report:

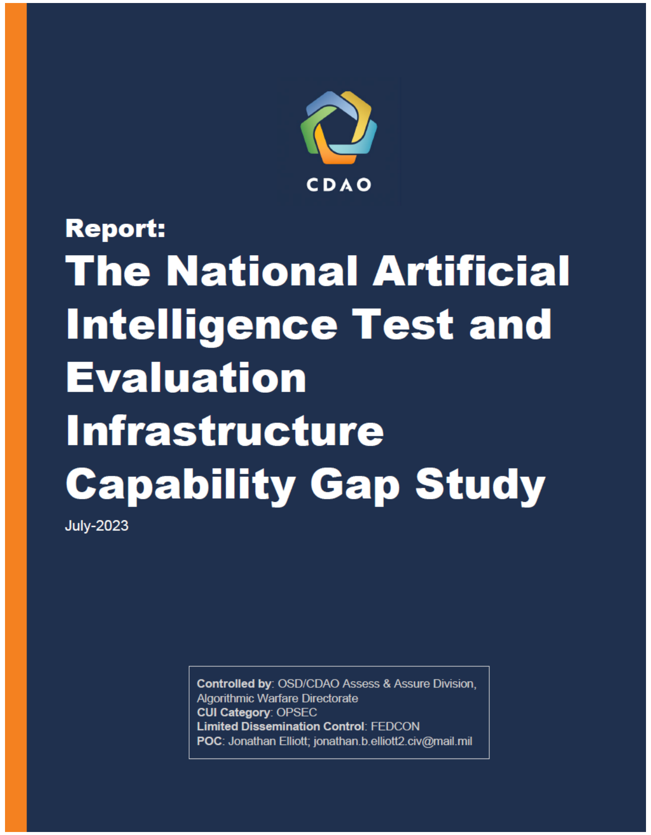
The National Artificial Intelligence Test and Evaluation Infrastructure Capability Gap Study

July-2023

Controlled by: OSD/CDAO Assess & Assure Division, Algorithmic Warfare Directorate
CUI Category: OPSEC
Limited Dissemination Control: FEDCON
POC: Jonathan Elliott; jonathan.b.elliott2.civ@mail.mil

What are the key problems?

1. Lack of maturity and domain knowledge in DoD AI testers
2. Difficulty in scaling tools across various DoD environments, platforms, and missions
3. Lack of tools for operationally-realistic conditions



CDAO

Report:

The National Artificial Intelligence Test and Evaluation Infrastructure Capability Gap Study

July-2023

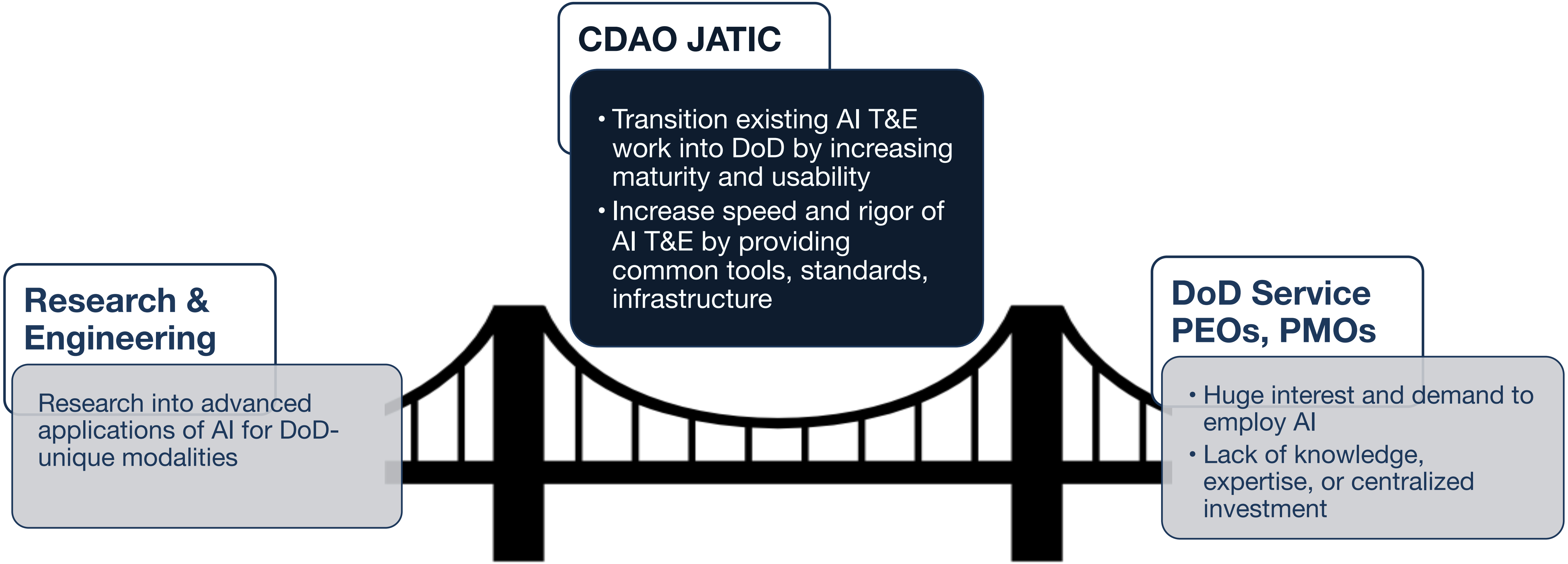
Controlled by: OSD/CDAO Assess & Assure Division, Algorithmic Warfare Directorate
CUI Category: OPSEC
Limited Dissemination Control: FEDCON
POC: Jonathan Elliott; jonathan.b.elliott2.civ@mail.mil

JATIC Scope

- We are focused entirely on **AI Model Testing**
- Why?
  - Applicability of tools across multiple missions and systems
  - Required domain knowledge for further stages of testing
- Within that, our initial focus is **CV Classification & Object Detection**

CDAO

Bridging the gap



CDAO JATIC

- Transition existing AI T&E work into DoD by increasing maturity and usability
- Increase speed and rigor of AI T&E by providing common tools, standards, infrastructure

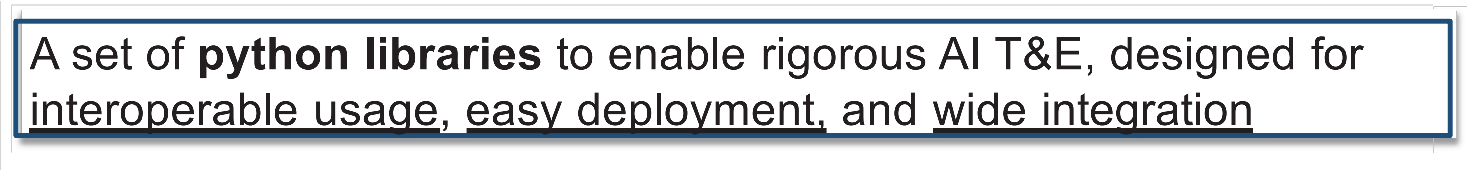
Research & Engineering

Research into advanced applications of AI for DoD-unique modalities

DoD Service PEOs, PMOs

- Huge interest and demand to employ AI
- Lack of knowledge, expertise, or centralized investment

AI T&E Libraries



A set of **python libraries** to enable rigorous AI T&E, designed for interoperable usage, easy deployment, and wide integration

- Straightforward deployment, setup, and use within variety of development or testing environments
- Seamless integration with key MLOps platforms and capabilities
- Using standardized model, data, and metrics protocols
  - Widely compatible
  - Easy-to-satisfy
  - Informative
  - Dependency-free

databricks

Amazon SageMaker

mlflow

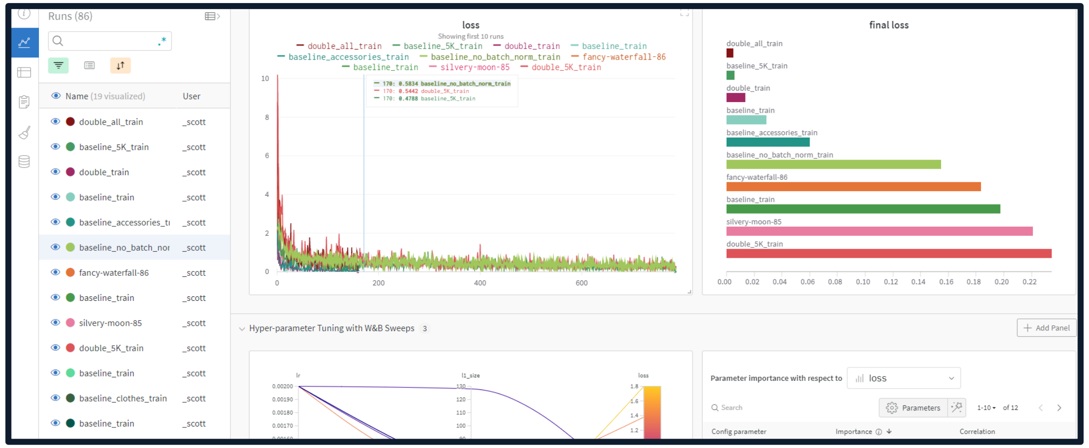
W&B

CDAO

T&E + MLOps

To be effective, AI T&E capabilities **must** *integrate seamlessly* with MLOps pipelines!

- *Continuous testing* of AI models **requires** this close integration. This is especially relevant as AI models must be retrained or fine-tuned more frequently
- Integration into MLOps provides incredible synergies between T&E and other AI/ML capabilities:
  - T&E + Workflow orchestration -> automated execution of model test plans
  - T&E + Model registries & experiment tracking -> improved T&E traceability and enhanced model metadata
  - T&E + Visualization dashboards -> seamless comparison between many models across test cases
  - T&E + Hyper-parameter optimization -> optimize model hyperparams for robustness, explainability, etc.
  - T&E + Labeling -> model T&E inference results inform potential errors in ground truth labels
  - …



Runs (86)

Name (19 visualized)	User
double_all_train	_scott
baseline_5K_train	_scott
double_train	_scott
baseline_train	_scott
baseline_accessories_tı	_scott
baseline_no_batch_nor	_scott
fancy-waterfall-86	_scott
baseline_train	_scott
silvery-moon-85	_scott
double_5K_train	_scott
baseline_train	_scott
baseline_clothes_train	_scott
baseline_train	_scott

CDAO

Capabilities

Tool	AI T&E Capability	Developer
Adversarial Robustness Toolbox*	State-of-the-art library of **adversarial attacks** and **defenses**	IBM
Armory*	Testbed for scalable evaluations of **adversarial attacks** and **defenses**	TwoSix Tech
Dataset analysis metrics library	Evaluate datasets for similarity, drift, and complexity	ARiA
XAI Toolkit*	Generate **visual saliency maps** on AI predictions using black-box and white-box techniques	Kitware
Natural Robustness Toolkit	Generate **operationally realistic data perturbations** and **augmentations** in-silico using **sensor-model** based techniques to test model robustness	Kitware
jatic toolbox	A source of common types, protocols, and utilities to enable synergistic and streamlined AI T&E workflows	MIT
rAI Toolbox*	Generate **data perturbations** and **augmentations** in-silico to test model robustness	MIT
Nebari*	Open-source AI & data science platform, designed for collaboration, scalability, and rapid deployment	Quansight
Terminus	Split dataset into training, validation, and test sets, without **bias across population subclasses**	MORSE Corp
RealLabel	Using model inferences, identify potential **ground label errors** within data	MORSE Corp
Gradient	Develop standard **AI T&E reports** in Powerpoint, directly from python	MORSE Corp

***indicates existing open-source capability**

RAVEN - AI T&E Platform

RAVEN is an orchestrated MLOps solution composed of open-source capabilities, specifically tailored for AI/ML testing

- JATIC python libraries are ideal for organizations who have *already adopted* an enterprise MLOps platform, such Databricks or Sagemaker
- For those without infrastructure, the **RAVEN** provides best-of-breed open-source tools to **jumpstart AI T&E from Day 1**
  - Deployable quickly to commercial cloud, on-prem, local machines, or HPC using Infrastructure as Code

- RAVEN Provides capabilities for :
  - Workflow orchestration
  - Model registry, experiment tracking
  - Database / object store
  - Visualization dashboard
  - Jupyter lab / IDE
  - Multi-GPU resource management

Program Assessement

CDAO

T&E of AIEC Lessons Learned

Through providing test support to various AIEC programs/processes CDAOs Assessment and assurance team has a variety of lessons learned to share

- Data splitting between T&E and Training is critical and must be constantly adjusted – you will almost certainly get it wrong the first time, so save data for future use.
- Need to constantly iterate on algorithm and operational metrics. Your starting metrics will not be the correct metrics for the system.
- T&E does not just inform your fielding decision. It is critical feedback to prioritize data collection, labeling, and model development roadmaps.
- With generative AI evaluating the human-AI system as a unit has become even more vital to understanding operational performance.
- T&E scoping and answering, “how much test is enough?” is exacerbated by AI

Collaboration & Access

CDAO T&E is actively seeking key government partners leading AI/ML to:

- *Transition research or S&T technologies for AI T&E and AI Assurance*
- *Support developmental testing of AI technologies to be integrated and fielded into larger systems*
- *Understand your AI T&E requirements, building AI T&E tools within JATIC to support*
- *Obtain feedback from you to iterate and mature our capabilities*

Join at https://gitlab.jatic.net with a *.mil*, *.gov*, or FFRDC/UARC email to get access to our current tools!

Questions? Please reach out at: CDAO-AI-Test@groups.mail.mil

CDAO