CDAO

Test & Evaluation of AI Capabilities

Triumphs, Challenges, and Opportunities

Jon Elliott, Chief, Test & Evaluation, CDAO

CDAO Mission & Functions

Mission

Accelerate DoD’s adoption of data, analytics, and AI to generate decision advantage

Functions

1. **Lead and oversee** DoD’s strategy development and policy formulation for data, analytics, and AI
2. **Break down barriers** to data and AI adoption within DoD institutional processes (serving as a “demanding customer”)
3. **Create enabling digital infrastructure and services** that support Components’ development and deployment of data, analytics, AI, and digital-enabled solutions
4. **Selectively scale** proven digital and AI-enabled solutions for enterprise and joint use cases
5. **Surge digital services** for rapid response to crises and emergent challenges

CDAO 2022 Priorities

LOE #1
Policy, Strategy and Governance

- Develop data, analytics, & AI adoption strategy
- Revise institutional governance structure
- Implement data decrees

LOE #2
Enterprise Infrastructure & Services

- Provide joint enterprise capabilities
- Expand enterprise data repository
- Establish a responsible AI ecosystem
- Conduct data & AI talent and literacy campaign

LOE #3
Enterprise & Joint Mission Solutions

- Track 2022 NDS implementation
- Develop ADA CIP/COP pipeline
- Establish dynamic campaigning capability
- Assess JADC2 interoperability
- Support enterprise business health & audit

Why AI Assurance?

CDAO

- AI Assurance is an opportunity for asymmetric advantage in the AI arena.
- DoD does not need to be fastest—we should develop fast enough and better.
- Assurance process provides arguments and evidence to establish trustworthiness

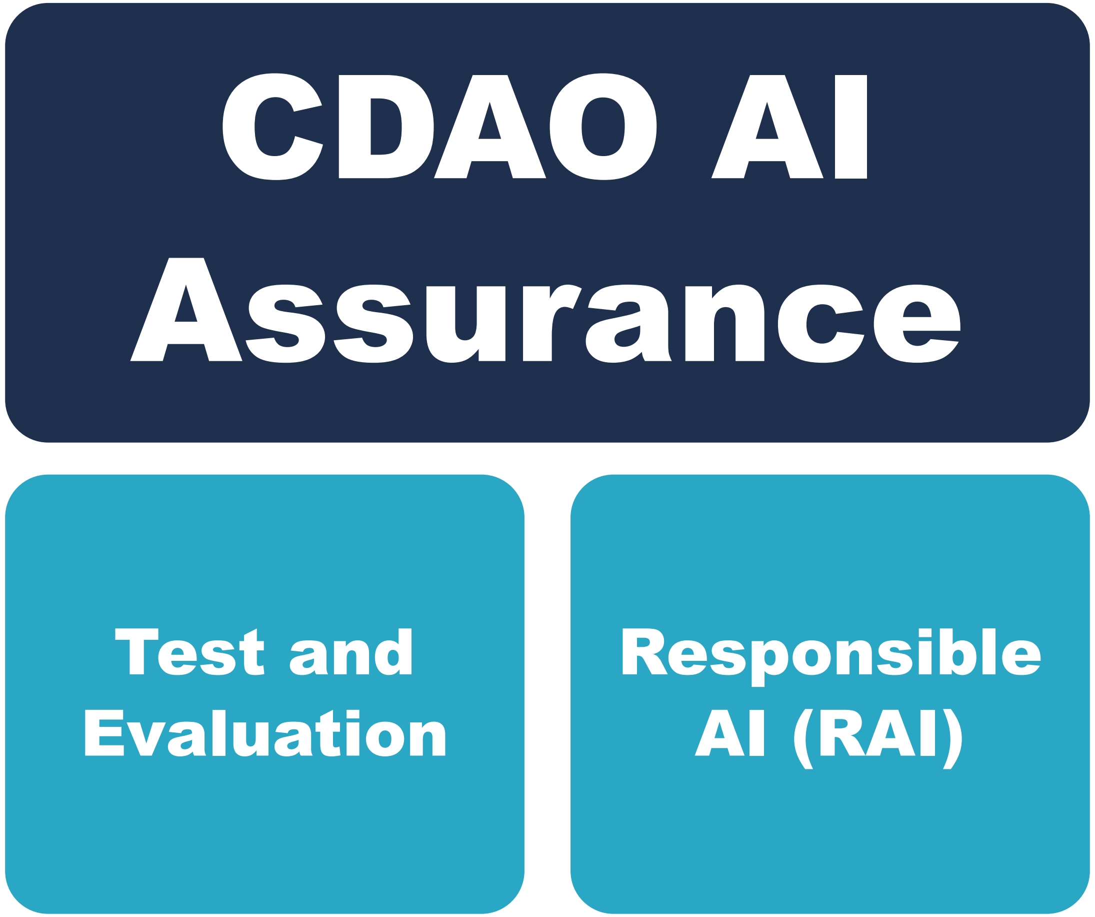
Arguments + Evidence = Justified Confidence

CDAO AI Assurance

AI Assurance Goal

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

Stakeholders include warfighter, commanders, PMs, acquisitions, regulators, tax payers, international allies



CDAO AI Assurance

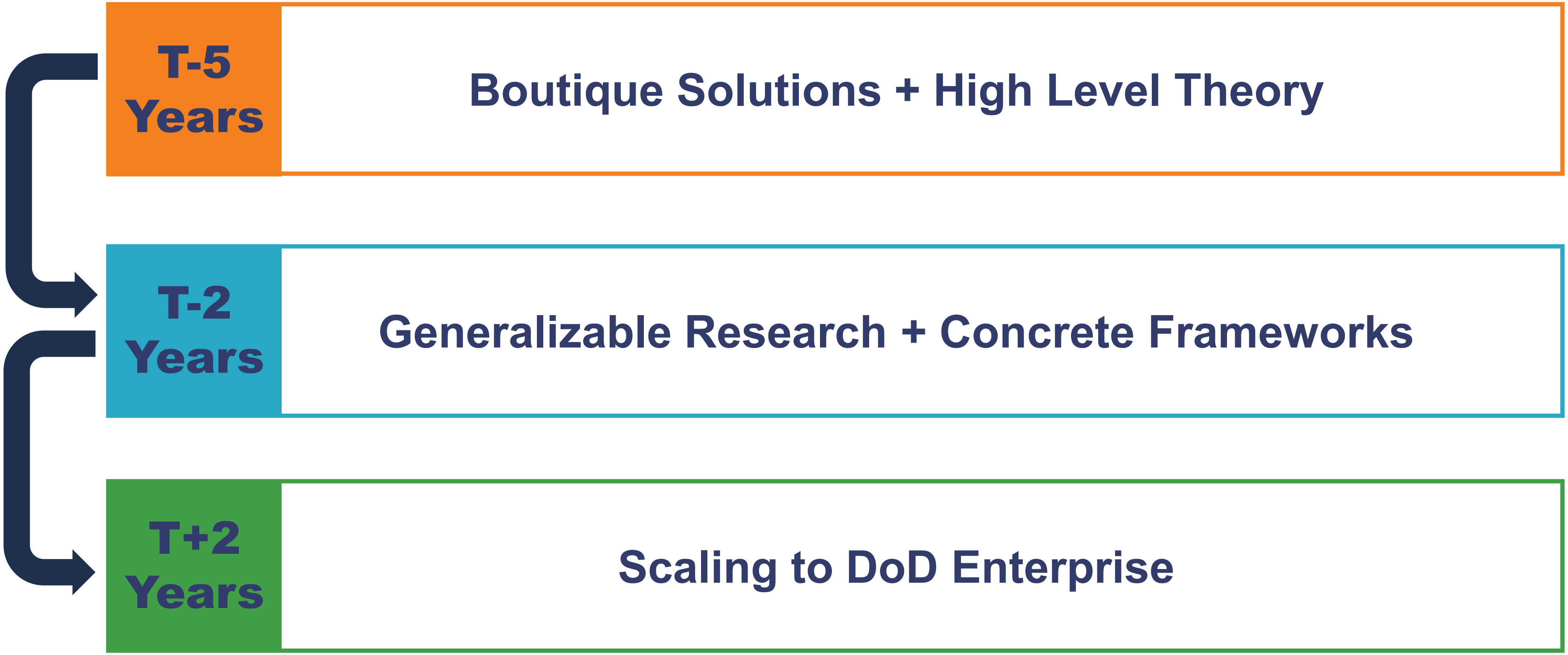
Test and Evaluation

Responsible AI (RAI)

DoD must update its T&E process for AI But we must produce our own solutions

- Tasks are dynamic and poorly constrained, success is hard to define/measure, and failures are catastrophic
- Can test only fraction of AI’s operational space but generalizing test results is difficult
- DoD must integrate adaptive T&E process across cradle-to-grave system lifecycle
- Systems upgrade and workforce reform are on the path to solutions

There is progress, but work remains



T-5 Years

Boutique Solutions + High Level Theory

T-2 Years

Generalizable Research + Concrete Frameworks

T+2 Years

Scaling to DoD Enterprise

Current Integration Efforts

CDAO

Research

Developing methods that will allow us to measure, model, and analyze AI system test results.

Frameworks

Guidance that will empower non-experts to create reasonable first passes at different aspects of test plans for these extremely complex systems.

Resources

Instrumentation, ranges, test beds, data, computational infrastructure, storage, and automated analyses that will enable effective T&E of AI.

CDAO

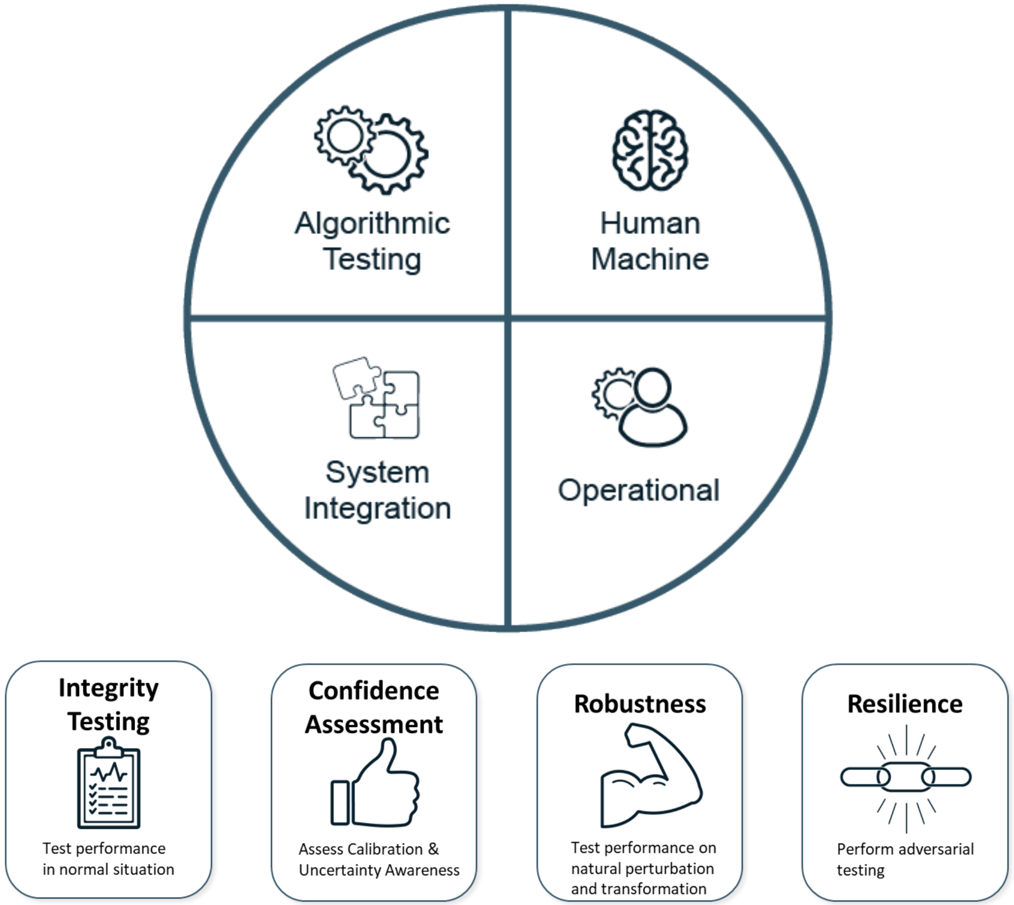
CDAO AI T&E Frameworks

Why Frameworks?

Help working-level testers without expertise in AI build justified confidence in DoD systems

- AIES T&E Framework
- Assurance Cases
- System Integration
- Human System Integration
- Operational & "Tactical" Testing
- Responsible AI T&E
- Trustworthy Autonomy – System Effectiveness

CDAO AI T&E Process



Algorithmic Testing

Human Machine

System Integration

Operational

Integrity Testing

Test performance in normal situation

Confidence Assessment

Assess Calibration & Uncertainty Awareness

Robustness

Test performance on natural perturbation and transformation

Resilience

Perform adversarial testing

Tailorable testing approach where each step is scaled to mission requirements

System Integration Test Framework

Functionality

- Ability of the system to do the work for which it was intended

Reliability

- Probability that a system performs correctly during a specific time duration.

Interoperability

- Ability of computerized systems to connect and communicate with one another readily

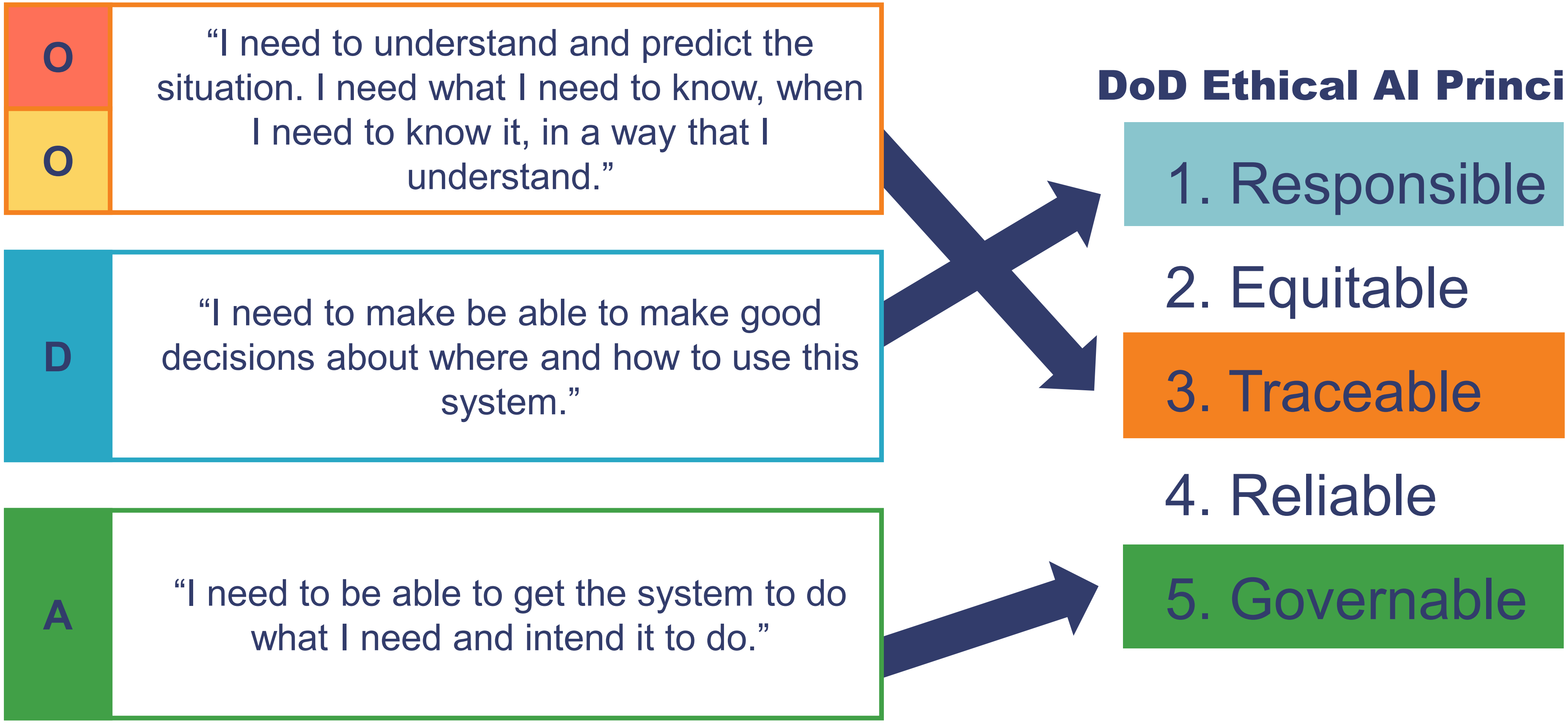
Compatibility

- Ability of multiple applications to interact in the same environment

Security

- Testing the systems vulnerability and weakness

Human System Integration Framework



O

O

“I need to understand and predict the situation. I need what I need to know, when I need to know it, in a way that I understand.”

D

“I need to make be able to make good decisions about where and how to use this system.”

A

“I need to be able to get the system to do what I need and intend it to do.”

DoD Ethical AI Principles

1. Responsible
2. Equitable
3. Traceable
4. Reliable
5. Governable

Human System Integration Framework

Observe & Orient

- Boundary Awareness
- Mental Model
  - Process Transparency
- Situational Awareness
  - Perceived vs. Actual
- Information Quality
  - Accuracy
  - Utility
    - Timeliness
    - Relevance
    - Completeness
  - Interpretability
    - Parsimony
    - Messaging Clarity
    - Appropriate Medium
      - …

Decide

- Trust / Reliance
  - Over-trust / Misuse
    - Complacency
  - Under-trust / Disuse
  - Envelope Understanding
- Emergence
  - “Off-label” Use
  - Tolerance for Misuse
  - Unpredictable Behavior
- Workload
  - Low Workload
    - Inattention
  - High Workload
    - Task Shedding
    - Invisible Work
- Judgement
  - Heuristics & Biases

Act

- Usability
  - Perceived Utility
  - Perceived Ease of Use
    - Affordances
    - Handling Qualities
  - User Intentions
- Behavioral Prediction
- Function Allocation
  - Oversight/Control
  - Handoff/Intervention
- Training, TTPs, CONOPS
- Middleware

Operational Test Framework

CDAO

Tactical Testing

- Smaller, but more frequent tests building towards IOT&E
- Multiple contexts and environments
- Variety of threats
- Stress individual aspects of the mission

Decision making is performance

- 3 types of decisions
  - Perceptual (“What?”)
  - Procedural (“How?”)
  - Executive (Goals setting)
- Influences how we test

Robust failure modes ensure safe testing

- Cannot test for everything
- Known unknown and unknown unknown failures will occur
- Need failures to be “graceful”

CDAO

CDAO AI T&E Research

RESEARCH: Methods

- Human Systems Integration (HSI)
- Sequential Test Design
- Learning Systems
- Explainable AI (XAI)
- Responsible AI (RAI)
- Adversarial / Red Teaming
- T&E in Deployment / Sustainment

T&E in Deployment / Sustainment

What is this effort?

A new addition to the CDAO AI Assurance portfolio

- How should systems be instrumented to enable continuous monitoring in fielded systems?
- What types of interventions are appropriate under what circumstances?

Why is this needed?

- Software will be continuously updated; possibly learning
- Data drift will degrade performance over time
- Adversaries will adapt to target system weaknesses

CDAO

CDAO AI T&E Resources

Joint AI Test Infrastructure Capability (JATIC) Vision

1. Interoperable set of state-of-the-art software for rigorous AI model & algorithm T&E which supports:

2. Many model architectures and deployment environments across the DoD

3. Across a wide range of AI operational tasks

JATIC Resources for AI T&E execution

Test Harnesses

T&E BPA

RAPT

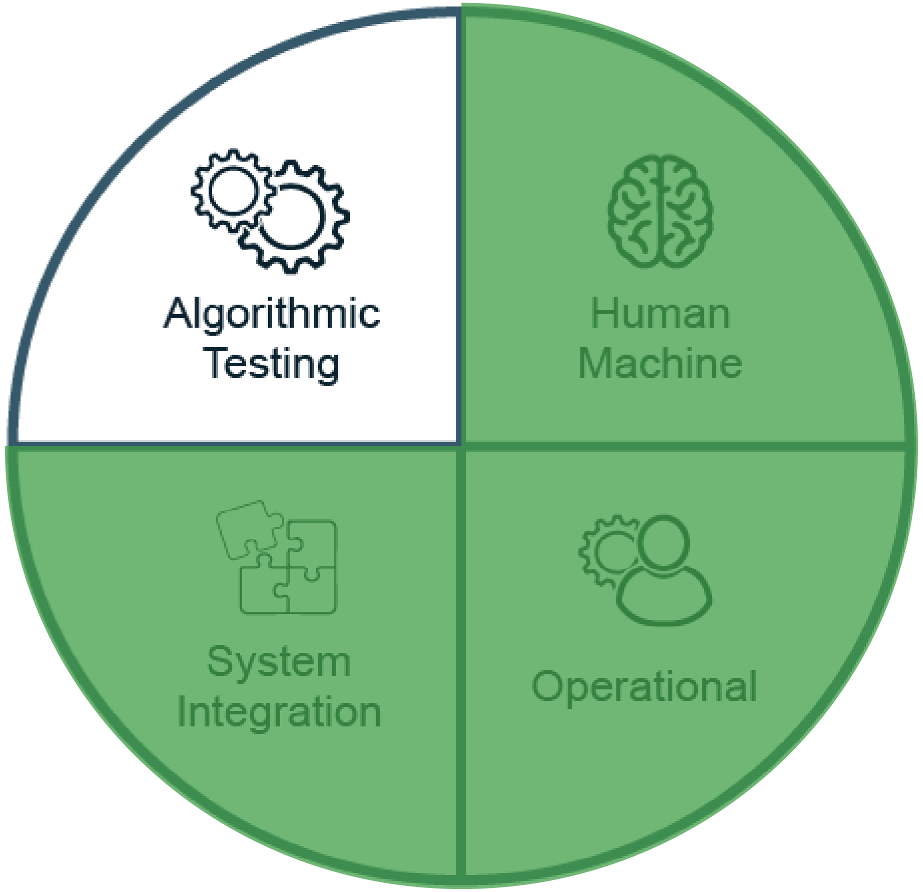
Test Data Capability

T&E Factory

T&E of Autonomy

T&E Factory Resources focus on Algorithmic Testing

- Model performance, reliability, and throughput testing
- Competence, confidence calibration testing
- Data perturbation and transformation
- Adversarial attacks and defenses
- Explainable and interpretable AI
- Model cybersecurity scanning
- Dataset distribution and bias analysis
- Label analysis and prioritization



Current T&E Factory Capability

DARPA GARD Armory

Armory is a platform for measuring AI model robustness to state-of-the-art adversarial attacks and efficacy of adversarial defenses, developed in the DARPA GARD program

Background

- Established theoretical ML foundations for adversarial attacks
- Encouraged the creation of effective defenses
- Armory combines the attacks and defenses into a comprehensive platform
- Collaborators included TwoSix, IBM, MITRE, University of Chicago, and Google Research

Armory design

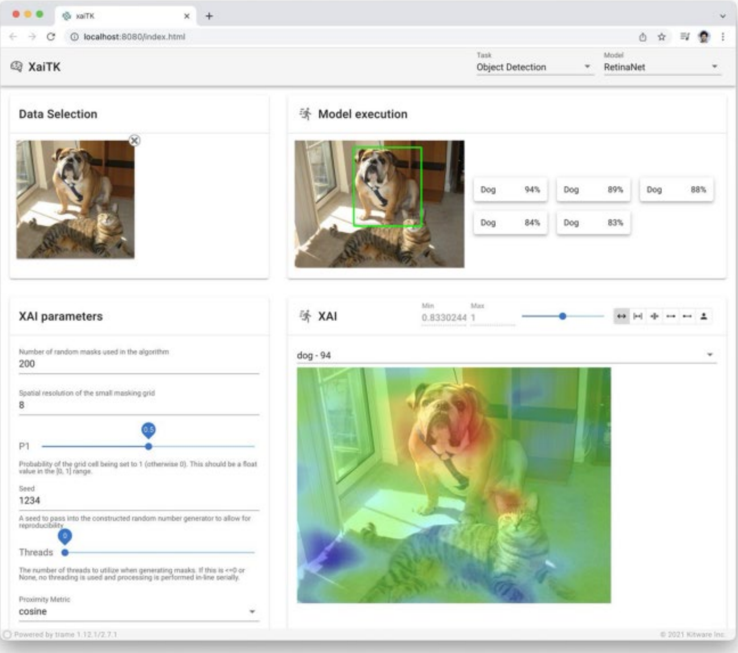
- **Extensible** – Supports continuous addition of cutting-edge attacks, defenses, scenarios, datasets, and metrics
- **Multi-domain** – Supports image, FMV, audio, and multimodal operational tasks
- **Realistic** – Evaluates models within realistic operational scenarios and adversarial threat models

https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception
https://github.com/twosixlabs/armory

Current T&E Factory Capability

DARPA XAI Toolkit

XAI Toolkit (XAITK) is a platform for visual saliency algorithms to enable of complex machine learning algorithms, developed in the DARPA XAI pr



Task: Object Detection
Model: RetinaNet

Dog 94%	Dog 89%
Dog 88%	Dog 84%
Dog 83%	

Min: 0.8330244
Max: 1

XAI parameters

Number of random masks used in the algorithm: 200

Spatial resolution of the small masking grid: 8

P1: 0.5
Probability of the grid cell being set to 1 (otherwise 0). This should be a float value in the [0, 1] range.

Seed: 1234
A seed to pass into the constructed random number generator to allow for reproducibility.

Threads: 0
The number of threads to utilize when generating masks. If this is <=0 or None, no threading is used and processing is performed in-line serially.

Proximity Metric: cosine

Background

- Created a suite of ML techniques to enable human users to understand, appropriately trust, and effectively manage AIECs
- Combines techniques and research into a unified and extensible package
- Collaborators included UC Berkeley, UCLA, Carnegie Mellon, UT Dallas, Texas A&M, Rutgers, Oregon State, SRI, Raytheon BBN, and IHMC

XAITK uses cases

- Establish justified confidence in AI models, addressing DoD Ethical AI Principles of Traceability and Reliability
- Enable feature understanding and analysis for black-box AI models
- Provide analysis of model failure modes and edge cases

https://www.darpa.mil/program/explainable-artificial-intelligence
https://github.com/XAITK/xaitk-saliency

kitware xaitk

Current T&E Factory Capability

Themis

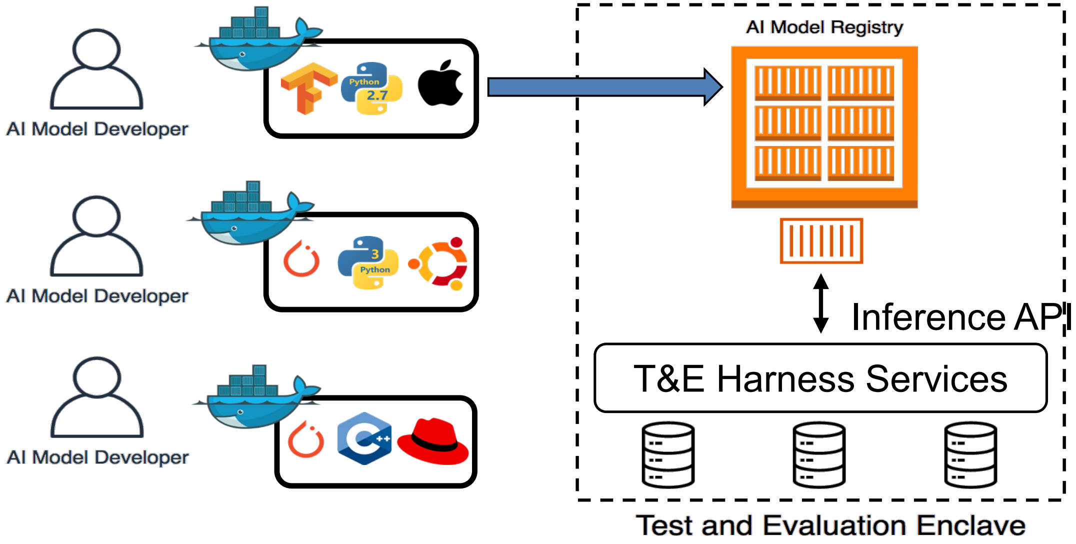
Themis is a test harness for rapid and repeatable T&E across a variety of AI models and deployment environments

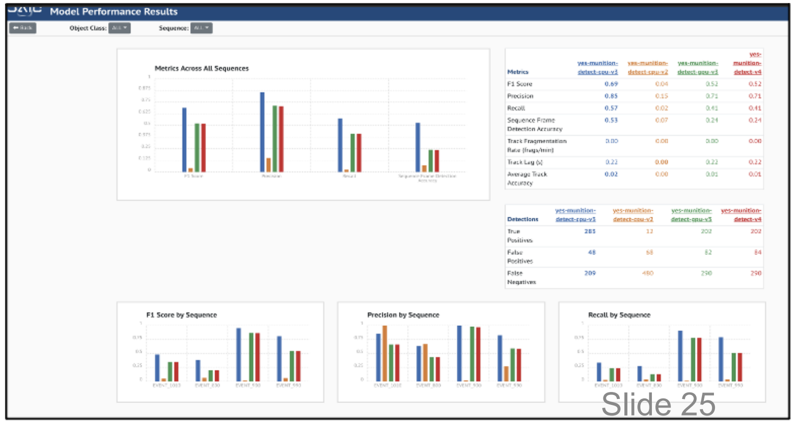
Background

- Developed by MORSE Corp to support CDAO T&E,
- Leveraged for FMV Object Detection and image segmentation AI T&E

Features

- Containerized microservices architecture enables portability and varied deployments
- Standardized "black box" model formats and APIs for wide interoperability across vendors
- Built with KeyCloak to enable vendor model submission and validation of test results




Model Performance Results

Object Class: ALL

Sequence: ALL

Metrics	yes-munition-detect-cpu-v1	yes-munition-detect-cpu-v2	yes-munition-detect-gpu-v3	yes-munition-detect-v4
F1 Score	0.69	0.04	0.52	0.52
Precision	0.85	0.15	0.71	0.71
Recall	0.57	0.02	0.41	0.41
Sequence Frame Detection Accuracy	0.53	0.07	0.24	0.24
Track Fragmentation Rate (frags/min)	0.00	0.00	0.00	0.00
Track Lag (s)	0.22	0.00	0.22	0.22
Average Track Accuracy	0.02	0.00	0.01	0.01

Detections	yes-munition-detect-cpu-v1	yes-munition-detect-cpu-v2	yes-munition-detect-gpu-v3	yes-munition-detect-v4
True Positives	285	12	202	202
False Positives	48	68	82	84
False Negatives	209	480	290	290

Current T&E Factory Capability

CDAO

VORTEX

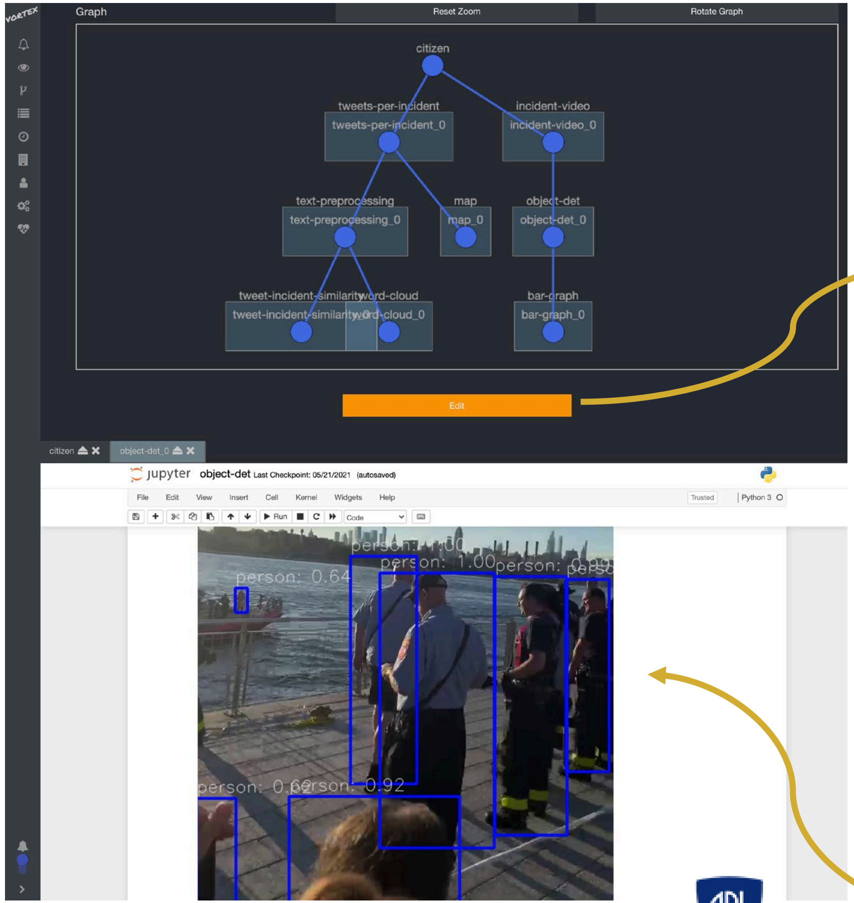
VORTEX is a scalable platform for streaming FMV AI model evaluation and data analysis

Background

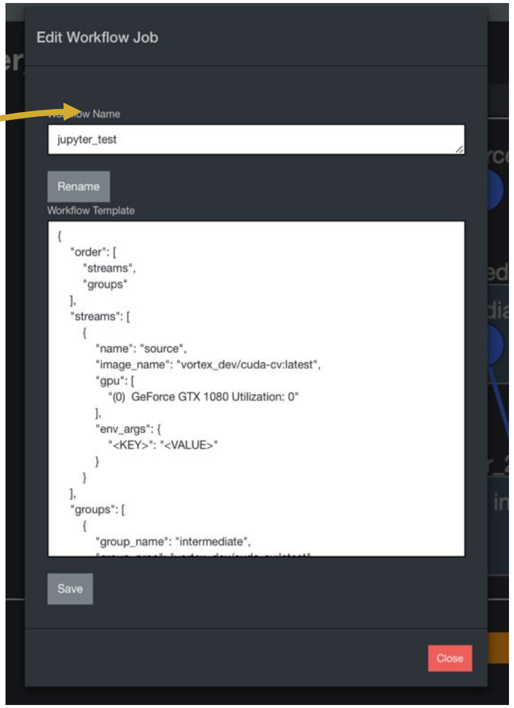
- Developed by JHU-APL, Vortex has supported T&E on multiple DoD FMV AI programs

Features

- Custom data processing pipelines that integrate modularly with other algorithms, such as data augmentation or model saliency
- Enterprise-wide availability, with containers on IronBank & Repo1



Dynamically modify workflow job



Edit Workflow Job

Workflow Name: jupyter_test

Rename

Workflow Template

```
{
  "order": [
    "streams",
    "groups"
  ],
  "streams": [
    {
      "name": "source",
      "image_name": "vortex_dev/cuda-cv:latest",
      "gpu": [
        "(0)  GeForce GTX 1080 Utilization: 0"
      ],
      "env_args": {
        "<KEY>": "<VALUE>"
      }
    }
  ],
  "groups": [
    {
      "group_name": "intermediate",
```

Save

Close

Implement custom workflow and functions within Jupyter notebook

Johns Hopkins
Applied Physics Laboratory

https://ironbank.dso.mil/repomap/parent?vendor=JHUAPL&product=VORTEX

Closing Thoughts

- Rigorous and timely T&E capability is a strategic advantage
- Science, policy, and infrastructure for AIES need to keep pace with technology development and modernization
- Department will require GPUs, CPUs, network stacks to perform super powerful, fast, low-cost AI
- We have made tremendous progress, especially in changing culture
- Much remains to be done!

CDAO

We need your help!

How are AI-enabled programs doing testing?

What metrics should be used to measure AI performance?

Do these tools provide value to you?