DEPARTMENT OF THE AIR FORCE
CHIEF DATA & AI OFFICE

# ARITIFICIAL INTELLIGENCE TEST AND EVALUATION PRIMER

FEB 2024

# Foreword

**Artificial Intelligence (AI) has forever transformed our military systems. Industry has quickly adopted this tool for commercial use, but the Department of Defense is responsible to the American people to ensure that AI-enabled military systems meet a much higher standard. To achieve that higher standard, these systems must be appropriately and efficiently tested and evaluated. While the world begins to comprehend AI's impact on our society, test professionals must answer the call to ensure our military applications are enabled and ready to perform and defend in our digital battlespace. Allow this Primer to be a first step for our test and evaluation professionals to understand and apply their expertise to Artificial Intelligence.**

- Col Tucker Hamilton, Department of the Air Force, Chief of AI Test and Operations

## EXPECTATION SETTING

This is a living document to provide T&E teams a resource to upskill and prepare for testing AI-enabled systems. Machine Learning (ML), a subset of AI, is currently disrupting the technology space and is the focus for this primer. Thus, uses of "AI" in this document may refer to AI in general or more specifically to ML (for more, see Understanding AI Technology). The primary audience is an individual that has a foundational understanding of both AI and T&E. Many topics are addressed with additional resources either hyperlinked or included in the appendices. This is not meant to be a checklist for testing AI because of the disparate nature of AI-applications. The Department of the Air Force (DAF) Chief Data and AI Office (CDAO) team will maintain this document. Please provide feedback to ADAx@us.af.mil. Reference to a contractor, award recipient, or any other non-government entity in this briefing is for informational purposes only and is not an endorsement.

It is crucial that all members engage their own learning journey about AI through community. In Appendix E you can find communities to ask questions and seek additional resources.

While this primer includes references to controlling statutory and policy provisions, it is not a formal policy document. Organizations should consult with their acquisitions professionals, legal counsel, or both before fielding AI capabilities or entering any agreement, to ensure proper adherence to laws and policies.

## ACKNOWLEDGEMENTS

We thank the following individuals and organizations for their contributions in the development of this document: Maj Joe Haggberg, Maj Riley Livermore, Mr. Adam Popovitz, and Lt Col Dan Riley. Interactions with Naval Test Wing Atlantic and Pacific Northwest National Labs were influential in determining the content of this document.

**LIST OF FIGURES**

## INTRODUCTION

Rapid advancements in useable Artificial Intelligence (AI) and Machine Learning (ML), a specific application of AI, throughout academia and the commercial sector require that the Department of Defense (DoD) adopt new Test and Evaluation (T&E) strategies. Fundamentally, AI is software. Software that is capable of learning, adapting, and responding to unforeseen conditions and environments and, therefore, requires an assurance posture akin to the continuous integration/continuous deployment (CI/CD) software lifecycle. While traditional DoD pathways for maturing, testing, and acquiring warfighting capabilities are not geared for the cyclical lifecycle of such AI-enabled capabilities, the Software Acquisition Pathway specified in DoDI 5000.02, Change 1, Sec 4.2 supports agile development methodologies and DevSecOps frameworks. Additionally, while AI is not synonymous with autonomy (see Test and Evaluation of Autonomy for Air Platforms), many autonomous systems will be AI-enabled or interact with AI systems. This requires T&E professionals to understand the software aspect of AI and the hardware testing requirements of autonomous physical systems. The principles outlined in this Primer are grounded in the substantial research done in the autonomy and AI T&E space.

## PURPOSE

This Primer is not a policy document but presents a framework for AI T&E compatible with the industry standard, Machine Learning Operations (MLOps), and the DoD's Operation of the Software Acquisition Pathway (DoDI 5000.87). Additionally, this Primer serves as a gateway to further resources that can be applied to the reader's specific AI T&E scenario. It is assumed that the reader has a solid foundational understanding of current T&E policies, practices, and deliverables for their respective organization and basic technical knowledge of AI. If the reader does not have a basic technical understanding of AI, the resources listed in Appendix B and Appendix C are a great starting point. Test teams tasked to test AI-enabled systems should ensure they have adequate AI education to make effective decisions related to test planning, execution, data analysis, and reporting for their specific project. More specifically, this Primer aims to:

1. Show where AI Test fits within DoD T&E
2. Outline the AI T&E Framework

## WHERE DOES AI TEST FIT WITHIN DOD TEST?

Testing AI presents an interesting paradox. It will require new tools, practices, and strategies to safely, securely, effectively, and efficiently test these systems. However, critical thinking, risk mitigation, and test fundamentals employed for decades will continue to play a central role in testing these systems. Figure 1 contextualizes AI testing within the current System.

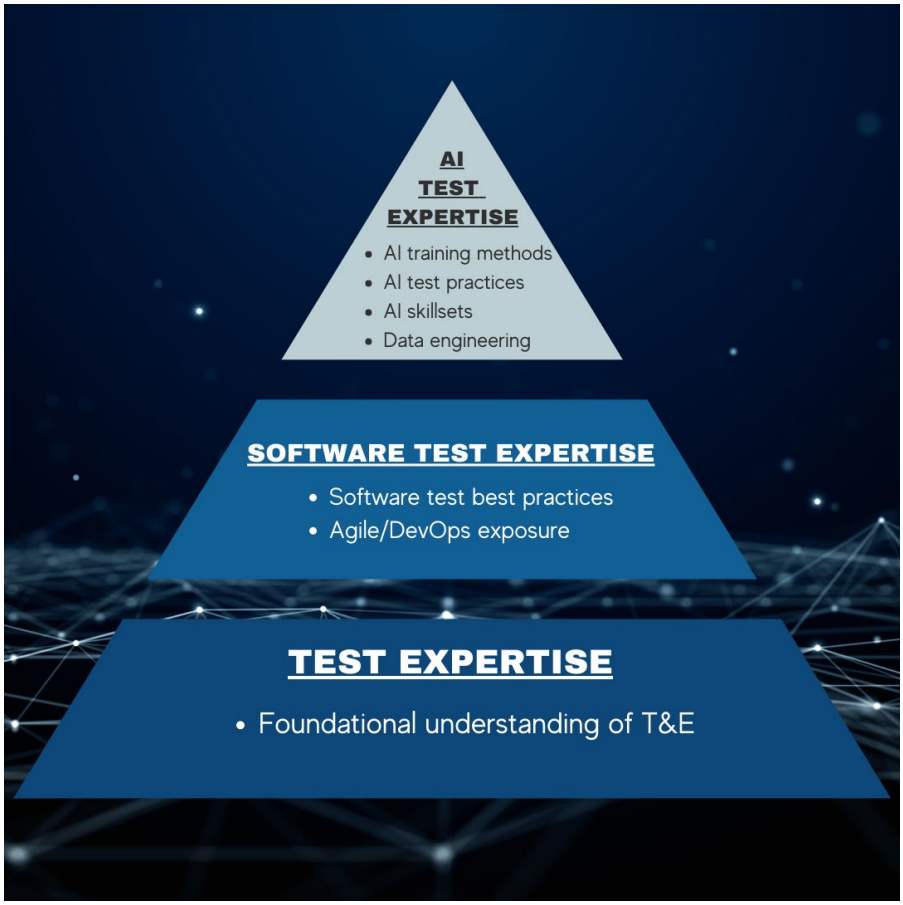


**Figure 1: Where AI Test Fits**

A foundational understanding and the core application of T&E principles continues to be invaluable. Software test expertise builds on this foundation. The success of MLOps should build upon the cultures of learning and collaboration core to Agile and DevOps because AI systems are fundamentally software. The level and depth of knowledge of software tests will depend on the nature of the AI-embedded System; however, some level of exposure to software tests should be considered for the T&E team. Testing AI-enabled systems will leverage this expertise, as well as demand new skills. Some of these skills include expanding data literacy, learning specific AI and ML training methods, test practices, skillsets, and an understanding of Data Engineering. Testing of AI-enabled systems will require true integrated test between developmental test (DT) and operational test (OT) organizations, and as such, this primer applies equally to all types of test.

## AI TEST AND EVALUATION FRAMEWORK

The AI T&E Framework, as seen in Figure 2, is an adaptation from the MLOps Process and Responsible AI Strategy's AI Product Lifecycle but is more specified to show T&E roles, considerations, and outcomes. The Framework, called the "4 Ds," outlines the phases of AI-enabled system acquisition: Discover, Design, Develop, and Deploy. The "Develop" and "Deploy" phases are where T&E traditionally occurs; however, the ultimate success of the AI-enabled System hinges on T&E professionals being involved from the beginning ("Discover" and "Design" phases). While the 4 Ds are shown to progress linearly, this process should be executed iteratively and incrementally. Moving through this AI T&E framework will depend largely on the acquisition

strategy, maturity of the AI technology, and the environment in which the AI-enabled System operates. Preeminent in the AI T&E Framework is the commitment to Responsible AI. The underlying six (6) foundational tenets of Responsible AI serve as guardrails during the four (4) phases of AI T&E.

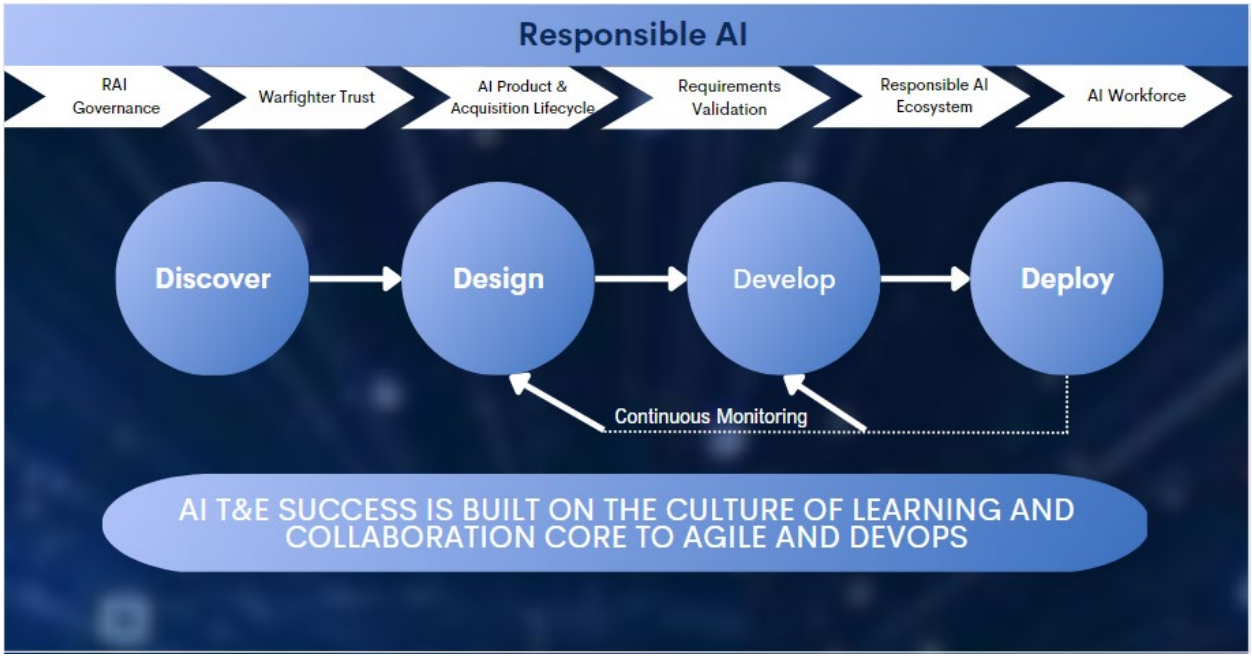


**Figure 2: AI T&E Framework**

RESPONSIBLE AI

In 2020, the DoD adopted the Ethical Principles for AI, encompassing the five major areas of Responsible, Equitable, Traceable, Reliable, and Governable. The Responsible AI (RAI) Strategy and Implementation Pathway provides six concrete LOEs to operationalize the ethical principles for AI (listed in Figure 2). One of the primary outcomes of implementing RAI is to achieve justified confidence in AI-enabled systems. The roles and responsibilities of AI governing agencies within the DoD are listed in Appendix G. Additionally, the National Institute for Science and Technology (NIST) AI Resource Center has provided the AI Risk Management Framework (AI RMF) to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems. As a consensus resource, the AI RMF was developed in an open, transparent, multidisciplinary, and multistakeholder manner over 18 months and in collaboration with more than 240 contributing organizations from private industry, academia, civil society, and government. Feedback received during the development of the AI RMF is publicly available on the NIST website. Responsible AI is not an aspect of AI T&E; it encompasses all of AI T&E and must be woven into each of the four (4) D's.

THE "4 DS" OF AI T&E

The "4 Ds" of the AI T&E Framework are described further in Figure 3. For each phase, three questions are answered:

1. What is the role of T&E?
2. What are the considerations that AI T&E professionals should have in this phase?

3. What are the AI T&E-specific outputs that feed into the subsequent phase?

The "Considerations" listed will undoubtedly have some component that applies in other phases. In addition, the specifics of the AI-enabled system acquisition and test program will further inform how to use this Framework.


**Figure 3: The 4 Ds of AI T&E**

DISCOVER

### 1.0 Role of T&E – Inform SMART Requirements

At the nexus of operators, developers, and program managers, test professionals have a unique opportunity to inform SMART (specific, measurable, achievable, realistic, and timely) requirements for AI-enabled systems. However, many program sponsors and offices may not be familiar with the intricacies of developing and testing AI-enabled systems. Therefore, the test team must take special care in addressing AI specifics from program inception. The DAF-MIT AI Accelerator's AI Acquisition Guidebook is a great first resource.

Given the stochastic nature of the AI models, the requirements must likewise have stochastic elements (i.e., full motion video (FMV) model must accurately identify vehicles >60% of the time when on feed for more than three frames. FMV Model should accurately identify vehicles 95% of the time when on feed for more than three frames). SMART requirements must tie outcomes to technical measures, metrics, and methodologies characterizing the risks of fielding the overall AI-enabled System. This process must also extend beyond technical performance to include OT concepts such as military utility and operational feasibility.

## 2.0 Considerations

The Discover phase is primarily the responsibility of a program office or other organization responsible for the acquisition of the program, but T&E professionals are key stakeholders in the process and can help to shape the program to success by being involved early.

### 2.1 Is an AI Solution Required?

Does the problem set require AI, automation, or other material solution? Automation is appropriate when the problem set is not dynamic (the variables do not change) and the answer is directly associated with known variables. AI is appropriate when the correct answer changes from moment to moment based on dynamic variables and there are sufficient data available for Training. Other material solutions may also be applicable based on the problem set and should be explored fully in the early stages of a program's life cycle.

### 2.2 Data Logistics

Data logistics refers to all the underlying infrastructure that supports an AI-enabled system. These logistics include security classification levels, access controls and management, data assurance, data types and formats, compute power, and storage requirements. Some good questions to ask for data logistics are:

- What data currently exist?
- How will the chain of data integrity be maintained?
- Who will be responsible for storing and managing the data?
- Who needs to access and to what type of data?

Wrestling with these questions will help inform what data logistics requirements will be needed for the successful execution of the AI-enabled System.

### 2.3 Hardware (H/W) Needs & Constraints

Often the hardware is the limiting factor of an AI-enabled system's performance, which makes identifying the H/W needs and constraints during the discovery phase paramount. To accurately account for the hardware needs, the whole lifecycle and ecosystem supporting the AI-enabled System should be considered. Some good questions to ask are:

- What are the desired payload form factors (size, weight, and power (SWaP) considerations) considerations?
- Where will the computing occur (Edge vs. Cloud)?

- What physical environments will the System operate in (i.e., weather, shock, and vibration considerations)?
- What test-specific hardware is required (i.e., surrogate platforms, sensors, hardware-in-the-loop facilities, high-performance computing)?
- How will the System receive and transmit data to other systems?
- What are other participating systems' hardware needs and constraints, and how will that impact the AI-enabled System's performance?

Additionally, it is essential to consider the supply chain and the country of origin as part of the hardware needs. Finally, the security and availability needs for the System will likely constrain the viable hardware options.

### 2.4 Software (S/W) Needs and Constraints

Analogous to the H/W needs and constraints, the software needs and constraints trade space should be explored in this phase. The s/w tools needed for data collection, visualization, and analysis must be considered depending on intended uses. Start by looking at the DAF's Data Fabric for various data tools, Platform One for app deployment, and Cloud One[1] for cloud architectures and services.

### 2.5 Cross-Functional Team

The roles of a team executing MLOps will look different than those executing in DevSecOps and include data-centric specialties. Identifying the core members as early as possible is essential for identifying all the requirements for the AI-enabled System. You can find an example of a list of different technical roles here. You should also consult with your servicing contracting and legal support teams early in the AI/ML development process. Additionally, this phase is a great time to identify any shortfalls in training and workforce development and start addressing those in preparation for the upcoming test campaign (ref Appendix B and Appendix C)

## 3.0 Output

### 3.1 Draft Requirements

These requirements will most likely be captured in an Initial Capabilities Document (ICD) or a Capabilities Needs Statement (CNS) per current DoD 5000.02 guidance. To the greatest extent possible, these requirements should be managed using Digital Engineering and Model-Based System Engineering (MBSE) best practices to pace the iterative and rapid nature of AI-enabled system design and development.

### 3.2 Program Standup

[1] A valid CAC is required to access

With the development of initial requirements documentation, a program of record should be established to shepherd and resource the AI system development and fielding.

DESIGN

## 1.0 Role of T&E – Build in Test Infrastructure

Given the complexity of AI-enabled systems, the test infrastructure to safely, securely, effectively, and efficiently test must be included during the design phase. This testing infrastructure can be considered tools and techniques and a built-in infrastructure for recording data (BIRD). Together, this infrastructure provides the primary means of creating a valid mental model for the System. Therefore, considering the tools, test techniques, and data recording capabilities in the Design phase is crucial for testers to ensure these AI-Enabled systems throughout their lifecycle credibly. The government test infrastructure may or may not be shared between the developer/contractor, but either way close coordination between the teams is essential to feed the correct testing requirements to the developer.

BIRD is fundamentally different than the "orange wire" used in the traditional DT efforts for two important reasons. The first is that, unlike the DT testing, the BIRD required for the AI-enabled systems will need to last that System's lifetime. Secondly, given how these AI models are trained and integrated, BIRD cannot be "bolted on" following development. Understanding the various considerations in this stage will help scope what BIRD looks like. Additionally, much of the data a government tester wants to see will be central to the developer's testing.

## 2.0 Considerations

### 2.1 Learning Rate

Understanding how the model will be trained and tested throughout its lifecycle is key to designing the appropriate T&E strategy and BIRD to assure the System. This learning rate can be seen as a continuum, with static and online learning models as bookends for periodic learning models.

- Static: Static Models do not change after deployment to the operational environment. Future versions may replace the existing model, but active Training at the operational site does not occur. As a result, static models have a lower sustainment cost but may be overly brittle and lack the robustness to operate effectively in environments for which it was not trained.
- Online Learning: These models are constantly updated based on real-world data and operations. Theoretically, these models are the most effective and robust because of the shortness of the feedback loop. There are two primary challenges with these models:
    1. The tested System does not represent the fielded System.
    2. Negative learning could be due to several things, including an alignment problem, erroneous human input, system noise, biased training data, or anomalous behavior being over-weighted.
- Periodic: Periodic is the hybrid approach that incorporates some of the benefits of both extremes while mitigating some risks. A periodic learning system can adapt to new environments and parameters while allowing the test community to provide some

assurance testing before fielding. The mission needs, the sophistication of the BIRD and test capabilities, and the end user's risk appetite will drive the learning's timescale.

### 2.2 Model Integration

AI-enabled systems are necessarily System of systems and will rely on numerous AI models to function safely and effectively. The two ways of approaching model integration are either modular or monolithic[2].

1. Modular: These modules are typically scoped to encompass a specific, well-defined, human-understandable task (i.e., one module for object detection and one for object classification). These modules are then integrated using a defined interface allowing data and information to be shared within the larger System. The benefit is that it's easier to isolate potential errors within the System because it's decomposed into human-understandable tasks (see compositional verification). In addition, the overall System is generally more interpretable, which fosters trust and a more accurate mental model. The downside is that modular systems carry more overhead for integration and are generally less effective than monolithic systems.
2. Monolithic: All of the functionality for a set of behaviors is incorporated into a single system (i.e., a system that detects and classifies an object). Monolithic systems require less training data and can generally yield better accuracy because monolithic systems are not constrained to operate in a human-understandable and intuitive way. However, this benefit is also a downside because they are less transparent which affects "trust" and when there are errors, they are much harder to isolate and rectify.

Additionally, identifying the model rights will be crucial to integration and testing. The level of visibility and insight that T&E professionals will have into the Systems needs to be defined early and will ultimately scope how well the integrated models can be tested.

### 2.3 Human System Interaction (HSI)

How a human user interacts with an AI-enabled system is crucial to understand and will drive how the System is tested and eventually fielded. It is also important to realize that the human and the AI relationship can change over time due to the System's maturity, operating environment, or integration with other systems. The different types of HSI are listed below:

- Human IN the loop: The human user is a vital link in the decision chain and owns the final decision. This construct is good for high-stakes applications of AI (i.e., autonomous kinetic effects), relatively low maturity, and early fielding of AI. However, the downside is that limitations often hamper the effectiveness of the AI system in human bandwidth and reaction time. Additionally, in complex, time-constrained environments, the user is prone to over-trusting the AI system, which could lead to undesirable effects.

[2] It is important to note that these two options are not mutually exclusive, and an AI-enabled system could have a large chunk of functionality that is monolithic with other modular components as well. Additionally, systems could start as monolithic and then transition a modular approach as it scales.

- Human ON the loop: The human user is not required to act for the System to function; instead, they monitor the system performance and can intervene in the event of an undesirable action (i.e., safety driver for autonomous vehicle). Not requiring human inputs to function mitigates some of the challenges caused by bandwidth and reaction time. This construct is best used for lower-stakes AI applications. Successfully implementing a human on-the-loop system is predicated on the quality of the human mental model of the AI and how well information is communicated to the human supervisor.
- Human INITIATED loop: The human is not in the loop at all and is only required to start the process by providing the objective to the AI. This construct is good because it is easily scalable. However, it should be only used when there is high confidence in the System or the stakes are sufficiently low because the alignment risk is always associated with the AI finding undesirable solutions to the human's objective.

### 2.4 Tools and Techniques

New tools and techniques must be created to test AI-enabled systems effectively. However, this does not mean that existing tools and approaches cannot be adapted as well (i.e., Design of Experiments or Formal Methods). You can find an example of a comprehensive literature review of tools and techniques geared for DoD T&E of Autonomy and AI is in Recommendation #2 of T&E of AI-enabled and Autonomous Systems: A Literature Review. Regardless of the tools and techniques selected, accommodations must be built into the AI-enabled System to facilitate proper use. Listed below are four general areas of focus that should be emphasized when designing the test strategy and corresponding BIRD:

- Automation: Building automated (and potentially even AI-enabled) planning, analysis, and reporting tools is essential to handle the size and complexity of AI-enabled systems at relevant speeds.
- Safety Middleware: A deterministic, rules-based approach that bounds behavior to reduce the likelihood of undesirable outcomes (sometimes implemented as Run Time Assurance). This approach helps mitigate safety risks for immature AI-enabled systems by allowing for limited capability fielding. However, safety middleware does carry its own development and sustainment burden, and care should be used before fielding a system with safety middleware to mitigate unintended cyber and tactical exploitation.
- Graded Autonomy: This principle is taken from the medical community and states that authority and autonomy should be incrementally applied, especially in complex and high-stakes endeavors. Once an AI-enabled system demonstrates justified confidence for a particular subset of its mission, the next increment can be tested. One of the techniques for achieving this is "shadow testing," which collects data on what an AI would have done, even though it doesn't have the authority to execute it. Successfully implementing shadow testing requires appropriate BIRD capabilities.
- Red-Teaming: The intentional attempt to exploit or degrade an AI-enabled system using various techniques (i.e., fuzzing, ablation, etc.). Known and suspected adversarial techniques should be incorporated into the test design.

### 2.5 Data Needs

The data type, quality, amount, and rights, as well as the corresponding data split strategy should be defined as early as possible. This includes not only the traditional training/test split used in development by contractors, but also additional test data for government validation. Ideally new test data will be collected during DT/OT, but it likely won't be enough for confidence in the system. It is also important to make sure the developer is not testing (and tuning) to the government test set. Additionally, any pertinent metadata need to be identified. Metadata in this context is information about the data, such as maximum file size, naming conventions, file creation time, etc. For example, an FMV model requires 200GB. The MPG4 dataset has a mix of urban, rural, desert, and jungle environments with visible cars. The maximum file size in this dataset will be no more than 100MB and have the following naming convention ENVIRONMENT_NUMBEROFCARS_DATEOFCREATION.MPG4.

### 2.6 Scalability

The supporting technology stack, integrations, data ingestion, and teams may differ between development and deployment. For example, in early development, a data fabric that supports model evaluation and iteration is more important than tools to monitor the model in production and manage network traffic (data latencies, user access, etc.). In other words, it is important to focus on integration within the data fabric rather than design around a specific tool or application. In addition, new development tools, data standards, or training techniques may come available, so the data fabric must be consistent with a good consistent with a good MLOps Strategy based on the implementation of MLOps Principles.

### 2.7 Cybersecurity

AI requires cyber protection because it is inherently software. Employing AI inevitably expands the potential cyber attack surface. The DoDI 5000.02, Software Acquisition Pathway includes considerations for cybersecurity in software. It is imperative to involve cyber testers and cybersecurity experts during the Design phase. Many of the coding packages and software being heavily leveraged in industry to manage data, perform analysis, as well as build and deploy the AI models (also known as the "Tool Chain") are open source models. Open source models are unlikely to be completely secure. Adversaries will look to exploit AI models in performing their intended function and impacts will include confidentiality, integrity, and availability. Examples of adversarial AI include input perturbation, data poisoning, model replication, and model inversion among others. Mitigating the most serious/probable cybersecurity threats should consider the time and diligence required to get an Authorization to Operate (ATO) for the particular cyber environment the AI and associated tool chain is being deployed on. The hardware (or IT systems) should also have consideration to the cyber threats posed by their supply chain. The DoD CIO website is the resource for finding guidance and information on topics related to cyber security including the Cyber Risk Management Framework (RMF) and Mission-based Risk Assessment for Cyber (MRAP-C).

## 3.0 Output

### 3.1 Requirements Documentation

Requirements could be a formal Capabilities Development Document (CDD) or a validated CNS. The requirements cannot just be technical specifications but also need to include operational performance metrics. T&E professionals should ensure that the test infrastructure

is captured in these requirements and that the requirements, in general, are testable with verifiable hypotheses. Lastly, these requirements should be structured with the most flexibility possible to account for changes based on model performance.

### 3.2 Test and Evaluation Strategy

Since the T&E phase will likely last the System's entire life cycle, this strategy should posture the test and operations community to think through those implications from the start.

DEVELOP

## 1.0 Role of T&E – Create a Mental Model

T&E professionals need to create a mental model to explore why an AI-enabled system produced the output it did. This mental model can be thought of in terms of model interpretability and explainability. You can find an example of both here. Creating an accurate mental model of the AI-enabled System is the first step in building a solid assurance case to understand when it can be expected to perform and when it cannot.

## 2.0 Considerations

### 2.1 Metrics and Evaluation Strategies

As mentioned, an AI model's effectiveness is determined by "what" it can do and "how" it operates in complex environments. At the model level, several well-defined metrics are used to assess model effectiveness (i.e., confusion matrix, root mean squared error, etc.). However, measuring and evaluating performance at the system level remains a challenge primarily due to the lack of accepted standards and criteria. Therefore, test professionals must consider how well the models behave and how the System behaves in its target environment. Using operational test (OT) concepts of measures of effectiveness and suitability (MOE/MOS) during the development and training of the AI models is crucial to ensuring the operational viability of the final product. For example, the Organization for Economic Cooperation Development (OEDC) maintains a repository of AI tools and metrics that can provide a great starting point.

### 2.2 Dataset Management

The test dataset should be appropriately managed to prevent leakage and ensure the AI models do not overfit the training data. In addition, it is essential to ensure these models are robust enough to operate within their operational context and that the evaluation metrics recorded during initial testing accurately represent the model's actual performance. The T&E team is responsible for articulating the data requirements to characterize the System effectively.

### 2.3 Modeling and Simulation (M&S)

Given the considerable state spaces of the AI-enabled systems and the desired timelines for fielding, the preponderance of the training and testing will occur using simulated data. M&S provides several significant benefits; some of the top ones are the ability to specifically test edge cases and to automate a large swath of the test matrix. Live-Virtual-Constructive (LVC)

capabilities extend the M&S environment to incorporate real-world entities, allowing for more representative testing. Effectively employed, LVC presents the best way to safely, securely, effectively, and efficiently bridge the gap between simulation and operational deployment of AI-enabled systems.

While M&S is essential, care must be taken not to overuse simulations when Training and testing these AI systems. No matter the fidelity of the simulator, there always exists a simulation-to-real-world (Sim2Real) difference that can impact the effectiveness of the AI-enabled System in the wild. In general, M&S capabilities can be split into two categories:

- Low Fidelity (LoFi) Simulation: Easier (and faster) to create and maintain. LoFi simulations also require less computing power and can be more easily integrated into LVC contexts. The challenge is deciding which facets of the simulation capability can be reduced. Often the fidelity threshold for M&S is not apparent and could result in ineffective or even damaging training data for the AI.
- High Fidelity (HiFi) Simulation: Much more realistic and more likely to generate useful training and testing data for an AI model. The main challenges are the increased computing power required and the time and energy to create and maintain the models and environments.

### 2.4 Model Interoperability

It is insufficient to ensure the AI models are in isolation and that the interactions between AI models that make up an AI-enabled system are equally important because it extends to the interactions between AI-enabled systems to ensure that undesirable emergent behavior does not occur. When these errors inevitably occur, placing all the blame on a single source is often difficult. Therefore, when interoperability issues arise, it is best to assume a maximum ownership mentality and patch all the interacting sources.

## 3.0 Output

There are several outputs in this phase that drive the test team to systematically build a mental model for the AI-enabled system:

- Test plans: Outline the objectives and data requirements to characterize the boundary conditions and edge cases that determine successful system performance.
- Evaluation cards: Provide metadata about the tools, processes, and data used to execute a specific test. These cards specify what tests are being conducted and what data needs to be collected in accordance with the test objectives outlined in the test plan. Evaluation cards should be detailed enough to allow for reproducibility of results. These cards can be seen as analogous to traditional flight test cards with specific modifications for testing AI-enabled systems.
- Model cards (example): Include metadata about the model that provides stakeholders with knowledge of the boundaries around a model's capabilities and limitations relevant and useful performance metrics. If given the training data, model cards should provide sufficient information to recreate the model.

- Data cards (example): Provide metadata about the data that the model was trained on, including label distributions, label subcategories, data quality, environmental factors, and distributions.

DEPLOY

### 1.0 Role of T&E – Sustain Assurance

T&E can expect to be most heavily involved in validating the AI-enable System before deployment. There may be cases where continuous monitoring and automation allow for the gradual easing of T&E involvement because the risks and assurance are being efficiently reported. If capability additions or integrations are included or expected in the AI-enabled System, the program may want to consider continual integration with the T&E team. Much like traditional programs, rigorous OT will be required to validate the system is ready for deployment, but more follow-on OT may be required throughout the lifecycle.

### 2.0 Considerations

#### 2.1 Continuous Improvement / Continuous Development (CI/CD) Pipeline

CI/CD is a core tenet of DevOps and reduces risk (both technical and programmatic) by making small and constant changes to software instead of batching several changes into longer intervals. As AI-enabled systems prepare for their initial deployment, the pipeline for CI/CD must be in place for patching and updating the System.

#### 2.2 Continuous Testing

Changes to an AI-enabled system's internal model should be considered equally to changes to the traditional software as a triggering event for new test events. The software code may not have any changes, but an updated model may have significant effects on performance that need to be identified.

#### 2.3 Data Logistics

Correcting the data logistics is foundational to supporting a CI/CD pipeline. Data logistics are again included in the Deploy phase because they'll have to evolve with the AI systems. Additionally, once the AI-enabled System is deployed "in the wild," it will be the first time the data logistics apparatus is thoroughly tested. This testing will likely require an iterative approach to support the AI system adequately.

### 3.0 Output

Outputs might look analogous to Deficiency Reports (DRs) / Watch Items (WITs) based on data collected from the fielded System. Still, they could also include qualitative comments from operational users identifying sources of frustration or system inadequacy.

CONCLUSION

The AI T&E Framework provides a roadmap for safely, securely, effectively, and efficiently testing and evaluating AI within the DoD. Anchored by Responsible AI, the 4 D's (Discover, Design, Develop, and Deploy) of AI T&E are based on industry (MLOps) and government

(Software Acquisition Pathway) best practices. In each of the 4 phases, T&E professionals serve an important role, which requires specific considerations and accompanying outputs. The ultimate success of the AI-enabled System hinges on T&E involvement from the beginning and that T&E infrastructure is baked into the System for its entire lifecycle. Furthermore, the AI T&E framework is malleable and must be adapted to the particulars of the AI-enable System's T&E requirements. Finally, as AI technology evolves, T&E professionals must remain educated and use new tools and technology to improve these systems.

## Appendix A - **ACRONYMS AND ABBREVIATIONS**

NIST AI Resource Center Glossary

| Abbreviation | Definition |
|---|---|
| AFIT | Air Force Institute of Technology |
| AFOTEC | Air Force Operational Test and Evaluation Center |
| AFRL | Air Force Research Lab |
| AI | Artificial Intelligence |
| AIA | Artificial Intelligence Accelerator |
| AIDA | Acquisition in the Digital Age |
| AIES | Artificial Intelligence-Enabled System |
| ANT | Autonomy and Navigation Technology |
| API | Application Programming Interface |
| ASC II | American Standard Code for Information Interchange |
| AISUM | AI for Small Unit Maneuver |
| AST | Algorithm Stress Testing |
| AUC | Area Under Curve |
| C2 | Command and Control |
| CDAO | Chief Data and Artificial Intelligence Office |
| CNN | Convolutional Neural Network |
| CONOPS | Concept of Operations |
| COTS | Commercial Off the Shelf |
| CSV | Comma Separated Values |
| CTF | Combined Test Force |
| CVT | Controls Verification Test |
| DAF | Department of Air Force |
| DARPA | Defense Advanced Research Projects Agency |
| DCG, | Discounted Cumulative Gain |
| DEO | Developmental Evaluation Objects |
| DIU | Defense Innovation Unit |
| DNN | Deep Neural Network |
| DOD | Department of Defense |
| DT | Developmental Testing |
| ECP | Entry Control Points |
| ETO | Executing Test Organization |
| FAA | Federal Aviation Administration |
| FAR | Federal Acquisition Regulations |
| FFRDC | Federally Funded Research and Development Center |
| GAN | Generative adversarial network |
| GAO | Government Accountability Office |
| GOTS, | Government Off the Shelf |
| GPS | Global Positioning System |
| HIL | Hardware In the Loop |
| HVI | Highly Valued Individual |
| IA | Information Assurance |
| IAW | In Accordance With |

| | |
|---|---|
| IEEE | Institute of Electrical and Electronics Engineers |
| INS | Inertial Navigation System |
| ITSC | International Conference on Intelligent Transportation Systems |
| ITT | Integrated Test Team |
| JAIC | Joint Artificial Intelligence Center |
| KPI | Key Performance Indicators |
| KPP | Key Performance Parameters |
| LL | Lincoln Laboratory |
| MAE | Mean Absolute Error |
| MIAB | Mag-In-A-Box |
| MIT | Massachusetts Institute of Technology |
| ML | Machine Learning |
| MRR, | Mean Reciprocal Rank |
| MSE, | Mean Square Error |
| MTBCF | Mean Time Between Critical Failure |
| NAIIO | National Artificial Intelligence Initiative Office |
| NDCG | Normalized Discounted Cumulative Gain |
| NGA | National Geospatial Intelligence Agency |
| NSW | Navy Special Warfare |
| NTPS | Naval Test Pilot School |
| NTWL | Naval Test Wing Atlantic |
| OSD | Office of the Secretary of Defense |
| OT | Operational Test |
| OTA | Operational Test Authority |
| PMO | Project Management Office |
| PSNR | Peak Signal to Noise Ratio |
| PTO | Participating Test Organizations |
| RMF | Risk Management Framework |
| ROC | Receiver Operating Characteristic |
| SF | Security Forces |
| SIL | Software in the Loop |
| SOTA | State of the Air |
| SQL | Structured Query Language |
| SSIM | Structural Similarity Index Measure |
| SUT | System Under Test |
| TPS | Test Pilot School |
| TQD | Training-Quality Data |
| TTP | Tactics, Techniques, and Procedures |
| UARC | University-Affiliated Research Center |
| UAS | Unmanned Aerial System |
| UAV | Unmanned Aerial Vehicle |
| USAF | United States Air Force |
| USG | United States Government |
| USNTPS | United States Naval Test Pilot School |
| UX | User Experience |
| WS | Weapons System |
| XAI | Explainable AI |

## Appendix B - **WHERE TO START**

CDAO Understanding AI Technology

The test team will require different skill sets from different functional areas. The Department of Defense Artificial Intelligence Education Strategy of September 2020 provides learning journey recommendations for T&E engineers (Figure 19 on page 33). Although the Education Strategy is not policy and is not a one size fits all approach, it is a useful document to review when to see how T&E engineers could fit into the overall workforce and the knowledge, skills, and abilities they should acquire.

**Asynchronous online Training (see Appendix C)**

**Publications**

- The MIT AIA publication page contains more than 100 scholarly works by the MIT faculty, students, and DAF members supporting the AIA.
- 2021 GAO report on an Artificial Intelligence accountability framework: GAO-21-519SP
- National Security Commission on Artificial Intelligence https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf
- Relevant literature that is recommended to be reviewed by test teams is a concept paper from the European Union Aviation Safety Agency (EASA) published in April of 2021. First usable guidance for Level 1 machine learning applications – Issue 01

## Appendix C - **WORKFORCE DEVELOPMENT**

Members of AI test teams typically fall into one of the following personas/archetypes:

- T&E Stakeholder: The stakeholder (or senior leader) is responsible for high-level decisions about a project, including the selection and deployment of AI models. This is often not one person and can include both program management and test leadership.
- T&E Engineer: The T&E engineer seeks to acquire unbiased information about the functioning of AI models in various dimensions to develop AI Assurance cases.
- Data Scientist: The data scientist thoroughly analyzes an AI model's performance and a dataset's constitution.
- System Administrator: The system administrator is responsible for oversight of their cloud environment.
- Model Developer: The model developer develops the AI model evaluated by the T&E tools before deployment.
- Model end-user: The model end-user uses the AI model in operation. Often, they are assisted by the AI model to make an informed decision.

Search free courses or start learning pathway courses related to Data Literacy, Data Science, MLOps, and others that fit within your archetypes from the CDAO Education Strategy:

- New AFeLearning (Percipio)
  - Enroll in a Skillsoft Bootcamp
- AFIT Continuing Education Courses
- PlatformOne Resources
- Lincoln Lab (LLx)
  - An educational collaboration between the MIT Lincoln Laboratory Supercomputing Center and the MIT SuperCloud
  - An account is required to register for courses, but registration is free to US Citizens with a .mil email account
  - The AI Foundations course linked below provides an exceptional overview of AI and ML (~24-hour online courses)
  - Artificial Intelligence Foundations | LLSC-SuperCloud Online (edly.io)
- DAF's DigitalU and Udemy
  - The DAF has partnered with Udemy to provide many free courses to registered users.
  - Navigate to https://digitalu.udemy.com/ to register for an account and unlock hundreds of rich AI and ML online courses
  - Here are a few notable courses:
  - Machine Learning A-Z: (In-depth – long, hands-on with code templates)
  - The DevOps Essentials – The Handbook (intro into DevOps and stages)

    - Complete MLOps Bootcamp | From Zero to Hero in Python 2022 (Overview of AI/ML pipeline)
    - Udemy Executive Brief on ML (Business-oriented explanation of ML)
    - Udemy Executive Brief on AI (Business-oriented explanation of AI)
- Gladstone
    - AI Fundamentals
    - AI T&E Introductory Course
    - Courses have a fee, and are available in ETMS for building a requirement
- Coursera
    - This site features courses from over 200 leading universities and companies
    - They offer many free courses, paid projects, certifications, and even degrees.
    - Visit their website for more information: https://www.coursera.org/
- edX
    - This site features courses from a 160-member consortium of higher education institutions (i.e., MIT, Harvard, Berkely)
    - Most courses are free, though you can pay for a certificate
    - Visit their website for more information: https://www.edx.org/
- O'Reilly Texts, Courses, and Events (free for those with a DOD ID)
    - Accessible through Continuing Education menu
- Miscellaneous
    - Commander's Read Ahead for AI
    - AI 101 Video (on MilTube)
    - Machine learning explained | MIT Sloan
    - https://www.ai.gov/strategic-pillars/advancing-trustworthy-ai/
    - Defense Acquisition University (DAU)

## Appendix D **- POLICIES/GUIDANCE**

DOD Software Acquisition Pathway

NIST AI Resource Center Technical and Policy Documents

CDAO Reference Library

DoD Chief Information Officer – AI and Data Strategies

DODI 5000.89 Test and Evaluation

Test and Evaluation of Autonomy for Air Platforms

DAU Test and Evaluation Enterprise Guidebook

## Appendix E - **OFFICES/WHO'S DOING WHAT**

- OSD Chief Digital and Artificial Intelligence Office (OSD CDAO) provides
  - Access to the Joint Common Foundation
  - Technical Assistance on AI programs
  - Testing and Evaluation Support
  - AI Product Lines
  - AI Policy and Ethics Guidance
  - Visit their website for more information: https://www.ai.mil/
- Defense Innovation Unit (DIU)
  - Accelerate DoD adoption of commercial technology
  - Transform military capacity and capabilities
  - Strengthen the national security innovation base
  - Visit their website for more information: https://www.diu.mil/
- Dept. of the Air Force – Chief Data and Artificial Intelligence Office (DAF CDAO)
- Dept. of the Air Force - MIT Artificial Intelligence Accelerator (AIA)
  - Strategic partnership between MIT and the Department of the Air Force
  - Designed to make fundamental advances in artificial intelligence to improve Department of the Air Force operations while also addressing broader societal needs
  - Research involves interdisciplinary teams, including Airmen, who collaborate across disparate fields of AI to create new algorithms, technologies, and solutions
  - Visit their website for more information: https://aia.mit.edu/
- Red Force AI
  - https://chat.redforceai.us/ (requires P1 account, CAC-accessible)
- Army Artificial Intelligence Integration Center
  - Partnered with Carnegie Mellon University
  - Connects with the broader artificial intelligence community
  - Basic research, force application, infrastructures & platforms, force integration, command & coordination, net-centric capabilities, sustainment, corporate management & support, battlespace awareness, and protection
  - Visit their website for more information: https://armyfuturescommand.com/ai2c/
- U.S. Naval Research Laboratory
  - Basic and applied research in artificial intelligence, cognitive science, autonomy, and human-centered computing
  - Adaptive systems, intelligent systems, interactive systems, perceptual systems
  - Visit their website for more information: https://www.nrl.navy.mil/itd/aic/
- MITRE Acquisition in the Digital Age (AIDA)
  - Transforming the federal acquisition environment using digital strategies and tools

- Their website has many features to help PMOs think differently about acquisitions
- Visit their website for more information: https://aida.mitre.org/demystifying-dod/

- Federally Funded Research and Development Centers (FFRDCs) and University-Affiliated Research Center Laboratories (UARCs)
  - Visit their website for more information: https://defenseinnovationmarketplace.dtic.mil/ffrdcs-uarcs/ or here https://rt.cto.mil/ffrdc-uarc/

- Assured DevSecOps of Autonomous and AI Systems (ADAS): TRMC's effort to be agile to the needs of AI and Autonomous Systems testers by removing barriers and strict processes to go after valuable technologies at technical readiness levels 1 through 8 from industry and academia.
  - AI Hubs – An OUSD R&E effort to build collaborative defense centers of AI capabilities around specific domains such as Computer Vision and Maneuvering and Reasoning. TRMC will run them and leverage the communities there to understand T&E resource gaps better.
  - Data Edge Teams – A deployable team of personnel and hardware to a test range or site to collect data from a test event and quickly and securely provide analysis and labeling before providing physical backhaul to a larger data center.

- Autonomy and Artificial Intelligence Test (AAIT): TRMC's effort to identify AI and Autonomy T&E gaps, develop AI and Autonomy T&E tools, and support maturation of needed capabilities up to technical readiness level (TRL) 6. AAIT maintains a wiki with all of their tools and their POC is thompsonm@mitre.org.
  - BAST – tool for recognizing when a system is no longer operating in the domain it was trained in and determining the minimal amount of testing to re-validate the System in the new domain
  - MIMIC – tool for surrogate training using imitation learning
  - RIOT – tool for finding internal bugs in node-based autonomous systems and determining if and how system-level inputs can exercise those bugs
  - STAR – automatically creates and executes test cases (TCs) to ensure and repeatable document that quantifiable robustness testing objectives have been met while providing testers with important testing details
  - SAFE – performs safety case evaluation on a system-level model
  - ICE-T – finds test cases in unlabeled video data where existing model performance is brittle

- Joint AI Test Infrastructure Capability (JATIC): OSD CDAO's program of record is creating a set of interoperable tools for AI algorithm T&E to be made available across the DoD. To be used in a wide variety of ML pipelines throughout the DoD, JATIC will emphasize compatibility, ease of use, and integration in its design.
  - The JATIC MVP will focus on techniques for the rigorous evaluation of computer vision (CV) algorithms. CV algorithms often serve as a critical foundation for more complex AI functionalities, such as autonomous agents. They are also represented

across various mission use cases, including ATR, ISR, satellite imagery, and medical imagery, which often use similar underlying AI/ML model architectures.

- JATIC will include capabilities that address the following areas of AI model T&E:
- Dataset Analysis capabilities allow T&E stakeholders to understand held-out T&E datasets, including their quality and similarity to operational data. Functionalities will include the computation of:
  - Dataset quality (e.g., label errors, missing data)
  - Dataset sufficiency for a given test (e.g., number of samples, variation)
  - Biases, outliers, and anomalies in the dataset, which may be naturally occurring or intentionally inserted (i.e., data poisoning)
  - Comparison of two datasets (e.g., divergence of dataset distributions)
- Model Performance capabilities allow T&E stakeholders to assess how well an AI model performs across a labeled dataset. Functionalities will include:
  - A comprehensive set of well-established CV metrics (e.g., precision, recall, mean average precision)
  - Metrics to assess probability calibration, i.e., the reliability of the model's measure of predictive uncertainty (e.g., expected calibration error, entropy, reliability diagrams)
  - Metrics to assess fairness and bias in model output across the test dataset (e.g., statistical parity)
  - Detect trends in model performance across the entire dataset, such as distinct types of model errors
  - Identify clusters of the model input space based on model predictions, feature values, network activations, etc.
  - Identify potentially mislabeled ground truth data based on sets of model predictions
  - Determine under-tested or high-value regions of the input space (for the model under test) to inform future test data collection or labeling
- Model Computational Performance capabilities allow T&E stakeholders to measure an AI model's computational efficiency, resource usage, and scalability. Functionalities will include the computation of:
  - Model throughput, latency, resource usage, and scalability, as well as constraints on these properties
  - Model performance across different hardware configurations
  - Optimal batch sizes for model inference
- Natural Robustness capabilities enable T&E stakeholders to determine how natural corruptions in data can impact model performance. These corruptions emulate realistic noise encountered within the operational deployment environment. Functionalities will include:

    - Pre-sensor, environmental or physical corruptions (e.g., fog, snow, rain, changes in target shape or dimensions)
    - Sensory corruptions (e.g., out-of-focus, glare, blur)
    - Post-sensor, in-silico corruptions (e.g., Gaussian noise, digital compression)
  - Adversarial Robustness capabilities enable T&E stakeholders to assess how adversarial corruption on data inputs may impact model performance. Attacks will include:
    - White-box and black-box methods
    - Empirical and certified attacks
    - Mathematical (e.g., Lp norm-constrained) and physically realizable (e.g., patch-based) attack.

Existing efforts in the DoD and industry:

- Data edge teams are available via the Tradewinds contract to design test collection strategies.
- The Cloud Hybrid Edge-to-Enterprise Evaluation & Test Analysis Suite (CHEETAS) framework provides a standard tool suite for building evaluation infrastructure for disparate acquisition portfolios. Developed and supported by TRMC, CHEETAS is free to the test community and is currently in use at multiple locations throughout the test community. In addition, CHEETAS provides a common GOTS analytics framework that:
  - Enables existing analysts to conduct data science
  - Emphasizes user time spent on analysis rather than data gathering
  - Provides consistent access regardless of data location and/or amount
  - Promotes sharing & reuse of tools & techniques across the community
  - Implements the DoD Data Strategy for RDT&E
  - CHEETAS is used today by these AF resources: AFOTEC Det 6, 413 FLTS, B-1, B-52, B-52, F-35
- The TRMC Joint Mission Environment Test Capability (JMETC) is a corporate approach for linking distributed live, virtual, and constructive (LVC) test resources. JMETC is designed to support the acquisition community during system development, developmental testing, operational Testing, interoperability certification, and Net-Ready Key Performance Parameters (KPP) compliance testing in a customer-specific Joint Mission Environment (JME). It provides readily available connectivity to the Services' distributed test capabilities and simulations. JMETC also provides connectivity for testing resources in the Defense industry.
  - Ryan Norman, Lead, Joint Mission Environments and Test and Training Enabling Architecture (TENA) Software Development Activity (SDA) Director, contact jmetc-feedback@trmc.osd.mil.

- TRMC is developing and validating a common architecture and requisite software to integrate Testing, Training, simulation, and high-performance computing technologies, distributed across many facilities.
    - TENA Middleware is the high-performance, real-time, low-latency communication infrastructure used by range resource applications and tools during the execution of a range event.
    - TENA Object Model enables semantic interoperability among range resource applications by encoding all the information that needs to be communicated among those range applications. An object-oriented design encapsulates the range community-wide set of interface and protocol definitions.
    - TENA Repository contains all the relevant TENA information not specific to a given test or training event.
    - TENA Logical Range Data Archive stores and provides for the retrieval of all the persistent information associated with a test or training event.

- TRMC AAIT contact: Vernon Panei <vernon.f.panei.civ@us.navy.mil>
- TRMC ADAS contact: Scott Padgett <scott.m.padgett.ctr@mail.mil>
- OSD CDAO contact: Jon Elliott <jonathan.b.elliott2.civ@mail.mil>

Existing efforts in the DoD: Implementations across the Department are isolated and generally employ commercial tools.

- Advana provides a data management platform for the OSD business process automation effort. This approach employs data engineers who curate data feeds that are ingested and normalized to a common ontology/schema. Commercial cloud-based solutionsare available for model development. Advana is accredited for CUI, Secret, and JWICS.
- Project Maven's data management platform, developed by Palantir, provides key data science functionality for analysis. However, the proprietary data storage format has created vendor lock-in at a premium cost.
- Early JAIC projects (i.e., Gargoyle and Smart Sensor), in collaboration with the Algorithmic Warfare Provisional Program Activity Office (AWAPPO), established a formal data management pipeline that has amassed a significant volume of labeled full-motion video (FMV) and synthetic aperture radar (SAR).
    - Unfortunately, the current data management pipeline is human-centric and manually intensive. Data is captured and labeled using a standard Unix/Linux file system, processing scripts, and read-only file mounts.
    - The no-frills approach avoids vendor lock but provides little in the way of data provenance or version control.
    - The primary vendor is building a government-purpose rights data management platform, Joint Enterprise Data-management System (JEDS), to alleviate the burden of working with the raw data at the file level.

- TRMC has been exploring a Collection at the Tactical Edge Testbed (CATE-T) concept wherein data edge teams work with test event planners to engineer a data collection plan from disparate and isolated systems. The targeted data lake for this effort, Repository of AI and Autonomy Enterprise, has not yet been implemented. Perhaps the data edge team/CATE-T concept could feed an existing data lake/warehouse-like VAULT.
- Joint Common Foundation (JCF) was an early attempt at an enterprise ML environment enabling decentralized AI development and Testing with a low barrier to entry to speed the delivery of AI capabilities—Advana is absorbing it.
- Existing efforts in Industry: For real-world data, most organizations adopt AI-assisted labeling to reduce fatigue on human users while increasing the volume and accuracy of labeled datasets. To further reduce the reliance on human labeling, academia, and industry are exploring the adoption of increasingly realistic synthetic data from generative neural networks and simulated environments to create accurately labeled datasets.
    - Modern data management systems provide a continuous feedback process wherein labeled data trains an AI labeler, which then pre-labels new data for human review. In this way, human guidance improves the AI labeler, which more accurately labels data
    - Highly realistic generative models have been created for specific objects/classes that can be tapped as a source of synthetic data. Once a generative model has been trained on actual data, it can generate synthetic data nearly indistinguishable from the real thing. In addition to accurate labeling, synthetic data in training and testing sidesteps ethical questions around bias and privacy in traditional datasets.

# Appendix F - **AI GOVERNANCE**

The AI governance organizations, OSD Chief Digital and AI Office and DAF CDAO, have several responsibilities that support T&E test teams:

- Data: provide sufficient, operationally relevant, usable datasets that DoD AI developers and T&E professionals can find and access to develop and test AI-enabled capabilities.
- Tools: make interoperable tools and reusable measures available in accessible tool repositories across the DoD to develop, train, test, evaluate, manage, and sustain AI-enabled capabilities across their lifecycle. Host standard T&E tools and datasets at an enterprise level to be available for rapid employment to support testing activities. These tools should strive to use open and interoperable standards that align with industry and facilitate the integration of programs to be tested.
- Users/operators in the lifecycle: involve users and operators early and often throughout the AI lifecycle to ensure the effective generation of use cases, requirements, T&E plans, human-machine teaming, and AI trustworthiness assurance methods. Ensure that there are processes for monitoring the performance of the AI-enabled capability as they are deployed in the field. This process should be able to trigger the reevaluation of the System.
- Interoperability: establish an AI T&E ecosystem across DoD that provides integrated and interoperable infrastructure, processes, and practices to streamline evaluations of effectiveness, safety, suitability, cybersecurity, and ethical employment. Ensure that tools and models adhere to interoperability standards set forth by leading industry standards.
- Workforce: develop a T&E workforce with AI technical expertise and is skilled in the tools, methods, and best practices for testing and evaluating AI-enabled capabilities.
- Adopt portable solutions that move seamlessly between development, testing, and deployment. For example, containerized solutions can be deployed to scalable cloud solutions during Training, development, and testing while they are evaluated against large numbers of variable conditions and situations. Ideally, such containerized solutions would be deployable as-is (i.e., without modification).
- Align with other DoD computing environments when possible. A federated and interoperable computing environment makes additional T&E data readily discoverable and usable across Services and, where possible, with Partner nations. Additionally, such alignment can minimize duplication of design and deployment efforts.
- Enable the secure transfer of data from the test ranges and operational environments to be used in developing accurate and functional tests.
- Coordinate with organizations that develop T&E tools to ensure that future DAF program T&E needs are represented.

Appendix G **- MISC**

As-A-Service Options Explained (e.g., PaaS, IaaS, etc.)

Cloud Security and Impact Levels Explained

Cloud Procurement Process

ML Ops Roles

DevOps at Google

AF Software DevOps

Agile Manifesto

Defense Innovation Board 2019 Software Acquisition and Practices (SWAP) Study