DOT&E, CDAO, DTE&A, TRMC, and Principal Director for Trusted AI/Autonomy Artificial Intelligence and Autonomy T&E Workshop

I. Introduction

On August 24th and 25th, the Office of the Director, Operational Test and Evaluation's (DOT&E) Chief Artificial Intelligence (AI) Officer, Dr. Kristen Alexander convened a workshop focused on the Test and Evaluation of Artificial Intelligence (AI) and Autonomous (AI-A) systems. The workshop aimed to initiate efforts under the 2023 DOT&E Implementation Plan for pioneering methods for T&E for weapon systems and other defense systems that are designed to change over time. The workshop was held in partnership with the Chief Digital and AI Office (CDAO) and three offices under the Undersecretary of Defense (USD) for Research and Engineering (R&E): their Principal Director for Trusted AI and Autonomy, Test Resource Management Center (TRMC) and Developmental Test, Evaluation and Assessment (DTE&A). The workshop was supported by Virginia Tech, the Institute for Defense Analyses (IDA), and Johns Hopkins University Applied Physics Lab (JHU APL). Nearly one hundred attended, who represent the breadth of the test and evaluation (T&E) enterprise across the Department of Defense (DOD). The gathering of these participants represented the establishment of a community of interest comprised of DOD Services, the Office of Secretary of Defense (OSD), non-profit organizations, academic institutions, and industry with shared interests and equities in advancing the definitions, documentation, and applications of T&E for AI-A systems.

Goals of the Workshop

To kick off the workshop, Dr. Alexander presented attendees with a challenge statement: AI/ Machine Learning (ML) present both a challenge and a potential benefit for the T&E community to harness. We must develop tools and processes to test AI enabled systems that allow us to determine the contextual and operational factors that influence operational effectiveness, suitability and responsible performance of AI/ML capabilities, especially as they learn and change during real operational use.

She identified two characteristics of a desired end state:

- Adequate assessment of operational and ethical performance of AI-enabled systems
- Adoption of AI and ML systems to make test and evaluation more effective, more efficient and more robust

The objectives of the workshop were to:

- Bring the **T&E community** together to discuss **unique considerations** of T&E and V&V of AI-A systems
- Identify and share previously identified frameworks. From that baseline, develop critical aspects **of T&E, V&V frameworks** for AI-enabled and autonomous systems
- Expand the **community of interest**
- Provide baseline understanding of **synergistic activities** across multiple organizations

The outcome of the workshop is this workshop report, which seeks to inform the path forward on frameworks, use cases, guidance, and investments.

Format of the Workshop

The format of the event began with baselining the current state of T&E for AI-A systems, provided as a series of presentations from DOD and research subject matter experts actively working to advance the policy, procedures, methodologies, tools, and skillsets in this field. AI-A T&E Frameworks were presented to provide framing and context to guide workshop discussions, along with future research, and events. An opening session laid out the key areas of emphasis for the workshop. Finally, the workshop moved into six breakout sessions with facilitated, interactive discussions around key questions and challenge areas relating to various T&E focus areas within the acquisition lifecycle for AI-A systems. The full agenda for the two-day "T&E of AI-A Systems Workshop" is provided below in Table 1.

Table 1. Workshop Agenda

<table>
<tr><th colspan="3">Day 1 - Thursday, August 24, 2023</th></tr>
<tr><th>TIME</th><th colspan="2">Description</th></tr>
<tr><td>0900</td><td colspan="2"><b>Welcome</b></td></tr>
<tr><td>0915</td><td colspan="2"><b>Baselining:</b> Overview of Current Frameworks and Synthesis
• CDAO / IDA, <i>Overview of AI/T&E Framework</i>
• CDAO / IDA, <i>National AI T&E Infrastructure Capability (NAITIC) Gap Study</i>
• CDAO / ARLIS, <i>Measuring Operational Impact in Combined Joint All-Domain Command and Control</i>
• US Air Force, <i>XQ-58 Flight Test</i>
• DTE&A / MITRE, <i>System Engineering Process for Testing AI Right (SEPTAR)</i>
• DOT&E / Virginia Tech, <i>Test, Evaluation, and Assurance for Learning Framework</i>
• DTE&A / STAT COE, <i>Autonomous Systems T&E Companion Guide</i></td></tr>
<tr><td>1115</td><td colspan="2"><b>Discussion:</b> Framework for Workshop</td></tr>
<tr><td rowspan="2">1300</td><td rowspan="2"><b>Day 1, AM Breakouts</b>
(select one)</td><td>T&E Preplanning</td></tr>
<tr><td>Model Development / Model T&E</td></tr>
<tr><td>1430</td><td colspan="2"><b>Break</b></td></tr>
<tr><td rowspan="2">1445</td><td rowspan="2"><b>Day 1, PM Breakouts</b>
(select one)</td><td>Live Virtual Construction T&E</td></tr>
<tr><td>System T&E</td></tr>
<tr><td>1615</td><td colspan="2"><b>Break</b></td></tr>
<tr><td>1630</td><td colspan="2"><b>Debrief Breakout Sessions</b></td></tr>
<tr><th colspan="3">Day 2 – Friday, August 25, 2023</th></tr>
<tr><td>0900</td><td colspan="2"><b>Welcome:</b> Recap Day 1 and Charge for Day 2</td></tr>
<tr><td rowspan="2">0930</td><td rowspan="2"><b>Day 2, AM Breakouts</b>
(select one)</td><td>Operational T&E</td></tr>
<tr><td>Sustainment & Model Updates T&E</td></tr>
<tr><td>1115</td><td colspan="2"><b>Debrief Breakout Session 3, Wrap Up, Next Steps</b></td></tr>
<tr><td>1200</td><td colspan="2"><b>Adjourn</b></td></tr>
</table>

This report details the findings, questions, and discussions that took place during the two-day event.

II. Baselining: Overview of Current Frameworks and Synthesis

Numerous organizations are currently working on solutions and methodologies for testing AI-A enabled systems. Those working on solutions span government, industry, Federally Funded Research and Development Centers (FFRDCs), and academia. As part of the workshop organization, the organizing team identified areas of promising research and application to share with workshop participants. Showcasing these efforts as a kickoff to the workshop provided a baseline for workshop participants to ground discussion. It also served to highlight organizations across the T&E community that are already working in this space. Baseline talks included subjects such as processes and frameworks, exemplars, and identification of gaps.

Rachel Haga from IDA presented the CDAO's AI T&E Framework and National AI T&E Infrastructure Capability (NAITIC) Gap Study. Carol Pomales from MITRE team spoke about System Engineering Process for Testing AI Right (SEPTAR), which was created in support of DTE&A. Charles "Charlie" Middleton from the Scientific Test and Analysis Techniques Center of Excellence (STAT COE) talked about the Autonomous Systems T&E Companion Guide that was also created in support of DTE&A. Major Ross Elder discussed XQ-58's flight testing through live, virtual, and constructive (LVC) activities. Dr. Joshua "Josh" Poore from the Applied Research Laboratory for Intelligence and Security (ARLIS) talked about Measuring Operational Impact in Combined Joint All-Domain Command and Control (CJADC2), which was in support of CDAO. Dr. Tyler Cody from Virginia Tech discussed a framework for test, evaluation, and assurance for learning systems. All the presenter briefings except for Dr. Poore's[1] are included as an attachment to this workshop report.

III. Workshop Focus Areas

The workshop was structured around six areas of focus that span the acquisition lifecycle. The focus of the workshop was clearly prioritized to not defining what AI or autonomy are, but rather focusing on how we adequately test AI-A enabled systems. Therefore, workshop participants assumed very broad conceptualizations of AI-A enabled systems, where AI was any "programmed ability to process information[2]" and autonomy was "the quality or state of being self-governing.[3]" Only six focus areas were selected due to time limitations and the need to target areas where questions regarding methods and processes need to be answered promptly. Figure 2 shows the focus areas and a general progression across the acquisition lifecycle. However, as was noted by many participants, focus areas are not distinct and progression is not unidirectional since as for learning systems, they often overlap and require iteration through many cycles.

[1] The ARLIS briefing is not approved for dissemination beyond the workshop.

[2] *John Launchbury, DARPA*

[3] *Merriam Webster Dictionary*

Acquisition Lifecycle Timeline					
T&E Preplanning: Scoping, Requirements, Acquisition Strategies	Model Development / Model T&E	Live Virtual Construction T&E	System T&E	Operational T&E	Sustainment & Model Updates T&E

Figure 1. Workshop Focus Areas

In an effort to ensure that the structure of the workshop did not result in the systematic exclusion of important topics due to the structure of the focus areas, workshop participants were asked to identify gaps in the workshops framing. One key area that participants emphasized were the need to engage operational users early and often, they mention terms such human machine teaming and human system interaction (50 total comments). This need to engage the operational user and consider human factors was reflected in the discussions under each of the six focus areas. Other areas that participants emphasized needed further attention included verification and validation (5 comments), data acquisition (2 comments), digital engineering (2 comments), risk management (2 comments), and systems of systems testing (2 comments). Figure 2 shows a word cloud of elements that workshop participants highlighted for future discussion.

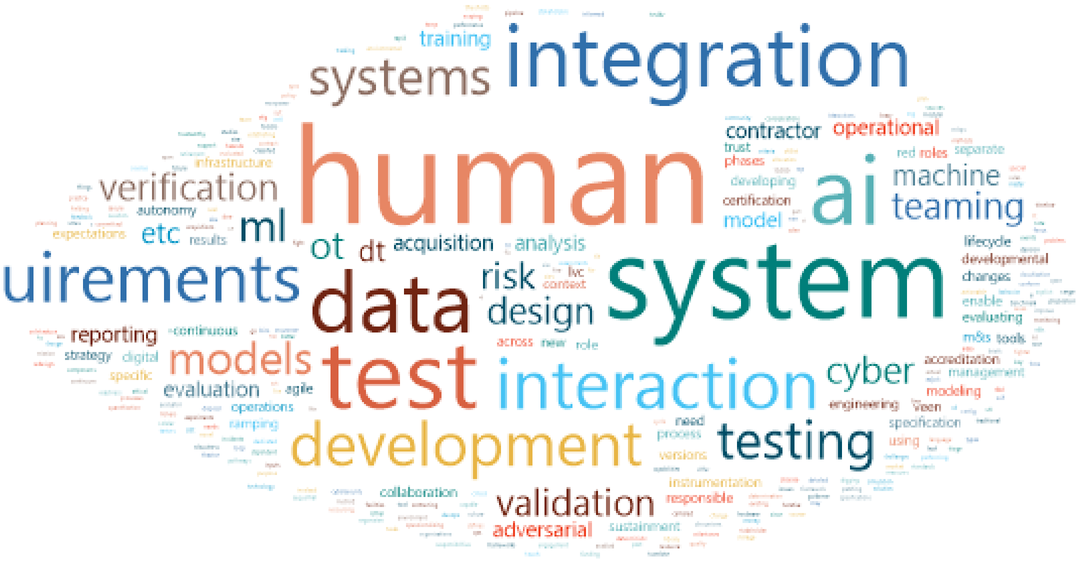


Figure 2. Word Cloud of Missing Elements

Workshop participants were also asked, in which focus areas do we expect to see the most changes across the acquisition pathways from today's practice? ***By far the most numerous comments were in the T&E preplanning section, highlighting the need for early engagement by the T&E community in the acquisition process to ensure we can test AI-A systems adequately***. The other area with a large number of comments was in the sustainment and updating focus area, where current T&E practices are not designed for systems built to change over time.

Finally, participants were asked to provide any additional resources that should be brought to bear by this community of interest. Those resources included additional talks, investments, test environments, education/training and more and are captured in Appendix A of this report.

IV. Breakout Sessions

Six interactive breakout sessions were held during the two-day workshop, in line with the six focus areas of the workshop AI-A T&E Framework. Sessions were conducted in three sets of two, with each session running for a duration of two hours. This setup allowed participants to join a total of three breakouts over the course of the workshop, with individuals self-selecting which breakouts to attend based on their interests and areas of expertise. The pairing and schedule for the breakout sessions is shown below in Figure 3.

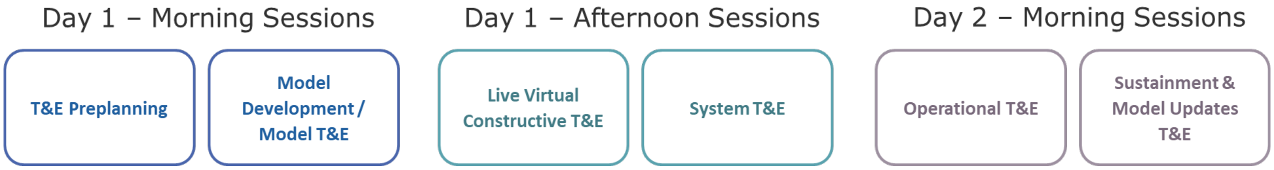

Day 1 – Morning Sessions
T&E Preplanning
Model Development / Model T&E
Day 1 – Afternoon Sessions
Live Virtual Constructive T&E
System T&E
Day 2 – Morning Sessions
Operational T&E
Sustainment & Model Updates T&E

Figure 3. Breakout Session Grouping

Each breakout session had its own pre-defined objectives and discussion topics, aimed at generating discussions on the unique challenges associated with the T&E of AI-A systems, along with ways to possibly address them moving forward. A summary of objectives and discussion topics is provided below in Table 2.

Table 2. Breakout sessions, Objectives and discussion topics.

Breakout Session	Objective(s)	Discussion Topics
T&E Pre-planning	• Explore the need for T&E engagement early in the acquisition process • Identify data challenges that are unique to AI-A T&E, and could impose significant risk to outcomes	• Scope • Requirements • Strategy and Plans • Engagement • Data
Model Development / Model T&E	• Explore the role of T&E in autonomous systems and AI model development	• Model Roles and Attributes • Metrics and T&E Results • Early System Development and Test Design Constraints
Live Virtual Constructive T&E	• Compare and Contrast System T&E vs Model T&E • Explore current strengths, weaknesses, opportunities, and threats for System T&E	• T&E Objectives • Test Designs • Data Characteristics • Resource Intensity
System T&E	• Explore the requirements for LVC to support T&E of AI enabled and autonomous systems	• Relationship between LVC and OT&E • Potential Impacts of LVC Environments • Need for Additional T&E Range Capabilities
Operational T&E	• Explore the progression from Model T&E → System T&E → OT&E for AI-A systems	• T&E Objectives • Measures and Metrics • Test Design and Methods • Data Pipeline
Sustainment and Model Updates T&E	• Explore the role of T&E is once an AI-A system is fielded	• Areas of Responsibility • Gaps in T&E

All breakout sessions were primarily facilitated by a T&E subject matter expert with experience in the use of human-centered design thinking to guide large-group discussions on topics relating to organizational changes and their impacts on policies, processes, and people. In an effort to encourage open and productive dialogue amongst participants, the facilitators asked all participants to follow three guiding principles:

1. Share ideas and talk openly during all breakouts, without concerns of attribution or retribution.
2. Enter into discussions with an "I believe" mentality, with the intent of having discussion focus more on exploring feasible ways to *overcome challenges and limitations*, as opposed to getting trapped by a mental roadblock of current limitations and constraints.
3. Begin discussions with the end in mind – with a shared goal of generating a workshop report that will form the basis for future frameworks and/or guidance.

Given the large size of the workshop and its sections, within each session participants were broken into smaller groups of 5-10 participants to facilitate discussion and generate a divergent set of ideas. The group members, discussion topics, and forms of engagement varied across the sessions. Throughout and at the end of sessions, facilitators brought the smaller groups back together for collective discussion and interaction between groups.

The remainder of this report section provides summaries of each of the six breakout sessions held during the AI-A T&E Workshop. Breakout session summaries include:

- Breakout objective(s) and discussion topics
- Key takeaways[4]
- Discussion summaries

Breakout Session #1: T&E Pre-planning

Breakout Objective(s) and Discussion Topics:

Objective 1: Explore the need for T&E engagement early in the acquisition process.

This objective examines the unique considerations relating to early engagement for the T&E pre-planning of AI-A systems. Breakout participants were initially given four topics for discussion, along with a set of associated questions, to explore unique considerations for AI-A systems and T&E roles and responsibilities:

- <u>Scope</u>. How do we establish (and measure) AI-A expected contributions for addressing known mission needs and operational gaps?
- <u>Requirements</u>. What does AI-A requirements development and management entail, taking into account the need for continuous testing of model and system performance?
- <u>Strategy and Plans</u>. Where should key elements of AI-A T&E planning be documented, and how should the document(s) be propagated across the lifecycle?
- <u>Engagement</u>. How does AI-A T&E planning require stakeholder engagement and SME inputs above and beyond the current T&E Working Integrated Product Team (WIPT) makeup?

Objective 2: Identify data challenges that are unique to AI-A T&E, and could impose significant risk to outcomes.

Breakout session participants were next presented with the list of AI-A T&E data challenge areas, and asked to define each challenge along with perceived risk to successful outcomes if they are not addressed. The participants were then asked to consider how enabling approaches may be used to mitigate the various challenges and perceived risks. Note: The definition of "successful outcomes" was left intentionally vague, so as not to hinder participant discussion. During discussion, participants tended to converge on outcomes focusing more on mission success, versus programmatic or technical success.

- <u>Challenge Areas</u>: Data capture, data rights, data quality, data protections and ownership

[4] The key takeaways were extracted by the report authors after reviewing workshop materials (pictures, sticky notes, flip charts) and notes captured by numerous notetakers during the workshop. The do reflect some interpretation of the data, but are provided to bring cohesion to the report.

- Enabling Approaches: Methods, tools, infrastructure, workforce

Key Takeaways:

An emerging theme from the T&E Preplanning Breakout participants was ***the importance of early and continuous engagement*** between AI/ML and engineering experts with operational and decision-making stakeholders.

While numerous rationales were presented on why this engagement is necessary, participants consistently emphasized three points:

(1) The need to incorporate mission and operational context in T&E planning across the lifecycle
(2) The importance of making sure human-systems interactions are adequately addressed during T&E
(3) The emphasis on having requirements that are flexible to accommodate continuous learning and advances in AI-A systems

In addition, engagement is necessary to help overcome several unique data challenges associated with T&E for AI-A systems, including:

- DOD is not accustomed to storing and turning realistic data into products for actionable, automated decision-making on various timelines and in different domains.
- Validating training and test data represent the intended operational environment is more complex and important for AI-A systems.
- The creation and management of a “golden data set” (a data set that is representative of and covers the full expected operational space that the system will be deployed to) is essential for AI-A system T&E. Such data should beseparate from training data, with access restricted so that model developers never have access enabling independent assessment. The data set must also evolve with operational mission changes.
- Transparency in data cards and data rights is essential.

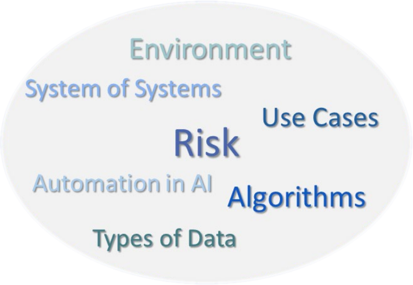
Discussion Summaries:

T&E pre-planning for AI-A systems requires unique process and procedural considerations early in the acquisition lifecycle compared with non-AI enabled systems. Today, acquisition programs often have limited engagement with end users and testers during the early lifecycle phases. The timing and levels of engagement need to increase for AI-A systems, but the rationale to support the “who, when, where, why, how” for these engagements is not well established. Additionally, resourcing will need to change in order to meet the need for additional engagements, which could mean increasing the T&E workforce, decreasing engagements or time on other T&E activities, or some other mix.

For this breakout, participants explored a total of five topics of importance for T&E pre-planning: scope, requirements, strategy and plans, engagement, and data. They discussed why each topic is important for T&E pre-planning; what early engagements would look like; unique challenges for AI-A systems; benefits and/or outcomes; etc.

The discussions challenged participants to think about T&E pre-planning in ways that differ somewhat significantly from today’s engagements. These differences are in terms of the need to plan T&E earlier in the lifecycle to address some of the unique performance assessment considerations of AI-A systems, and also in terms of a more general need for more continuous collaboration and coordination between T&E (including Service testers and OSD test authorities), operational user (warfighter), engineering, and acquisition program management communities.

Scope
The scope for AI-A systems should come from top-down derivation that connects Strategic Priorities → Mission and Operational Needs → Operational and System Requirements → Use Cases. Scope should be warfighter driven, and include operationally relevant use cases that inform T&E planning. Once the understanding of mission context and operational relevance is achieved, then programs and testers begin building requirements with performance metrics.



Risk management is increasingly important for AI-A systems, and in turn the use of T&E for risk assessment and monitoring is also increased.

Additional terms used frequently during the breakout session, that highlight considerations for future discussion: testable requirements, specification of algorithms, risk, and system of systems.

Requirements
AI-A systems necessitate requirements that are both *flexible* and *operationally focused.* In particular, low-level requirements specifying system performance must be flexible enough to accommodate the continuous learning of the AI-A system in dynamic environments, while still enabling rigorous T&E preplanning to determine what needs to be tested and how. In addition, metrics for measuring and assessing these requirements should conform with SMART criteria – specific, measurable, achievable, relevant, and time-bound – and should include well-defined acceptance criteria. This approach to defining AI-A system requirements allows for early planning and scoping of T&E to inform resourcing, infrastructure and instrumentation, data, safety, and continuous cybersecurity. These T&E planning considerations span the end-to-end lifecycle, including during sustainment until its disposal.

Breakout participants placed significant emphasis on keeping the operational user as the central user when defining requirements. Benchmarking was identified as a way to define what is “good” versus what is “good enough” for AI-A system performance, and should be used to generate stakeholder agreement on “acceptable risk” along with requirements prioritization and documentation.

Additional terms used frequently during the breakout session, that highlight considerations for future discussion: Continuous cybersecurity T&E, safety and risk mitigation.

Strategy and Plans
Discussion on strategy and plans revolved around the question of whether existing acquisition program artifacts are sufficient for documenting T&E plans for AI-A systems, or are new artifacts necessary. Participants strongly agreed that new artifacts are not required; instead, existing documentation may be expanded. In particular, Test Plans, T&E Master Plans (TEMPs), and Test Strategies.

In order to support the unique considerations for T&E of AI-A systems, Test Plans and TEMPs need to account for novel, new technologies, with documentation on AI-A T&E specific data plans, data cards, and model cards. In addition, requirements, metrics, and measures need to be explained and evaluated with evidence (perhaps with new approaches to design of experiments). In addition, Test Plans and TEMPs must become living digital artifacts, with a well-defined scope and operational environment, to account for continuous feedback from the field and other input sources on performance and requirements.

These unique considerations for test planning and documentation drive toward the eventual goal of building the assurance case(s)[5] for AI-A systems in critical areas of functionality.

Contracts and training data rights remain a challenge to be addressed. There was broad agreement from the participants that having more information about data sets and models used in AI-A would lead to more productive and efficient testing. However, participants noted that contracting makes it difficult to get information about data sets and models in many circumstances, and that they are not able to directly engage with DevSecOps or MLOps pipelines directly, a process which underlies many best practices from industry (more details in "Data" below).

Engagement

The single and most significant takeaway on "engagement" for T&E preplanning was: Tell your story, tell it often, tell it all over the place so the purpose and intent of the AI-A system is understood, accepted, and funded. All participants agreed that this is part of creating "market value" proposition and technical knowledge at all levels of engagement. Additionally, the participants also agreed that beginning engagement with discussions on responsible AI (RAI) goals is critical, but currently it is not known how to accomplish RAI or who is responsible.

Security Access
AI, ML, Data Science SMEs
SAP Squadron
Warfighter
Market Value Proposition
Up & Down Chain

Additional discussion on this topic reflected themes that are consistent with those from the "scope" and "requirements" topics. In particular, the warfighters are the most important stakeholder group for T&E engagement, with additional stakeholder groups including senior leaders, AI-A subject matter experts, and the public. Participants noted that inputs from these stakeholder groups on an iterative cadence provide valuable information to maturing AI-A projects, but it can be difficult to get their time when supporting testing is not a primary part of their job so it can be difficult to create engagement opportunities.

A recommendation was made to consider following the US Air Force model for establishing a Special Access Program (SAP) squadron for testing highly classified systems early in the lifecycle. The benefit is faster testing with testers who are already cleared.

Data

The data challenges relating to AI-A T&E are vast, and cover a wide range of topics that impact the entire machine learning operations (MLOps) Cycle. Challenge areas include, but are not limited to: data capture, data rights, data quality, data protections, and ownership/stewardship. Enabling approaches to

[5] An assurance case is defined as An argumentation pattern consists of a given claim (or conclusion), the associated evidence or sub-claims and the argument of why the claim could be concluded in a given context and/or given restrictions

address these challenges are defined broadly as methods, tools, infrastructure, and workforce, but the levels of detail on what could / should be provided by each enabler is not well characterized.

Discussions on this "data" topic challenged participants to think about data in a way that stretched the imagination of "what if" and "then what". They considered how risks associated with data and the T&E of AI-A systems compare to risks with more traditional weapon and defense systems. And further, how risk mitigation through enabling approaches such as infrastructure and workforce compare across system types.

One of the key takeaways from the discussions: DoD is not accustomed to turning data into useful outputs that support actionable, automated (or semi-automated) decision-making on various timelines and in different domains.

In addition, while there are some overlaps with the data challenges faced by AI-A systems and traditional systems, the challenges are much greater and more complex – as are the potential risks if left unmitigated. Identified challenges included:

- validating training data itself, which is an order of magnitude more complex and important for AI-A
- ensuring transparency in data cards, data ownership, and data rights
- the need for a "golden data set" for testing that is separate from training data. Here a "golden data set" included concepts such as government ownership, spanning the operational environment, including data labels and high quality metadata.

The approaches (or enablers) for addressing challenges are also more complex because the data associated with AI-A systems is significantly more resource intensive, and requires significant investments in new / novel infrastructure, workforce, tools, and methods. Participants attempted to identify where changes and/or improvements are needed in each of these areas, creating the list of considerations shown below in Table 3.

Table 3. Enablers to overcome data challenges associated with the T&E of AI-A systems.

Instrumentation	**Tools**
• Instrumentation • Storage • Access • Transferability • Realism • Fusion • Security classification	• Cross-domain solutions • Synthetic data generator • Expeditionary analysis tools • Test automation • Coverage • Preprocessing • Auto-labeling • User survey data • Software applications • Data poisoning • Model drift • Modeling & simulation
Workforce	**Methods**
• Training • Qualifications • Annotation quality • SAP team	• Contracting • Policy • Synthetic data • Tracking / usage • Accessibility • Formatting • Standards • Data management • STAT analysis • Training on false data

Breakout Session #2: Model Development and Model T&E

Breakout Objective(s) and Discussion Items:

Objective: Explore the role of T&E in autonomous systems and AI model development.

At the onset of the breakout session, the scope of discussion for model development and model T&E was established as the *AI-A models under test*, as opposed to the modeled environment in which the systems are measured. Participants were presented with four topics relating to model development, aimed at eliciting subject matter expert discussion on unique considerations and challenges. Below is the list of four topics:

- <u>Model Roles and Attributes</u>. Including AI algorithm under test, AI models used to test and measure other algorithms, those used to train AI, and the interoperable components of a full test environment that is made up of multiple overlapping AI models.
- <u>Metrics and T&E Results</u>. Including explainable AI, edge cases, and requirements.
- <u>Early System Development and Test Design Constraints</u>. Including data, test environments, relationship between design and burden of testing.

Key Takeaways:

Throughout the breakout session, Model Development and Model T&E breakout participants frequently emphasized ***the importance of establishing a common lexicon for AI-A systems*** that includes further specificity of terms like “model” to support common understanding and coordination between AI/ML experts, contract specialists, testers, operational community, and others. This is needed to be able to immediately (via shared terminology) differentiate between an AI algorithm, a trained AI model under test, AI models used to test and measure other models, and the interoperable components of a full test environment that is made up of multiple overlapping AI models.

Additional key takeaways included:

(1) T&E may be leveraged to derive numerous explanatory metrics across system performance, operational performance, and LVC testing.

(2) T&E informs and guides model development, and helps define the requirements for scaling AI capabilities in the field.

(3) Data, test environments, and determining “how much testing is enough” are pinnacle issues for T&E of AI-A systems.

Discussion Summaries:

Model Roles and Attributes

As mentioned above, discussion on model roles largely focused on the need for common lexicon and further specific of terms like “model”, to support common understanding and coordination between AI/ML experts, contract specialists, testers, operational community, and others. This is needed to be able to immediately (via shared terminology) differentiate between an AI model under test, AI models used to test and measure other models (potentially made without AI), AI components that may be composed of multiple AI models, and systems that may be composed of both AI components and non-AI components.

In addition to the breakout of these different model roles, thoughtful T&E should always categorize specified model attributes. Some characteristics include:

- Learning Style (e.g., reinforcement, supervised, unsupervised, or semi-supervised),
- Deployment State
- Domain (data type(s), potentially including multiple modalities),
- Function (inference, prediction, optimization, etc.), and
- Other model details (federated learning, transfer learning, online learning))

Designations by learning style, domain, function, learning parameters, and how a model will be used within the T&E domain will guide how best to test a given system.

The breakout team then discussed how the act of testing itself can help to derive numerous explanatory performance metrics across system performance, operational performance, and LVC testing.

Metrics and T&E Results

Participants recommended including the following parameters for metrics and T&E results:

- Sensitivity
- Model latency and throughput
- Boundary conditions
- Uncertainty quantification
- Epistemic uncertainty / confidence
- Accuracy and reliability
- Precision / recall bounds
- Feature significance
- Performance: confidence and stability
- Mission suitability
- Privacy
- Traceability and transparency
- Anomalous behaviors
- Decisions made by the AI
- Failure modes
- Robustness, resilience, and fragility

Understanding that these parameters will be recorded and measured, T&E informs and guides model development and helps define the requirements for scaling AI capabilities in the field in the following ways. First, it forces explainability and provides a feedback loop for AI model improvement. It identifies areas of concern as well as areas of peak performance, assisting with the design of edge cases. It highlights points of instrumentation and follows changes in metrics as models mature. And it offers different kinds of assessments – whether testing or validating concepts, scalability, or model limitations and breaking points. The T&E process helps define the requirements for scaling AI capabilities in the field, whether scaling models from single scenarios, integrating them with different models, or integrating them with outside systems.

Early System Development and Test Design Constraints.

With respect to early system development and test design considerations, breakout participants shared many challenges and areas for future growth within the practice regarding T&E development and model T&E. For example, the automated test of an AI model requires ground truth data, and different data are needed in different places throughout the process. The community needs to determine where data should or can come from as well as how and when to pull it. Case in point, reserve data may need to be pulled early and in larger than expected quantities for testing; however, this will be at odds with AI model developers who want to use all available data because it is usually associated with better performance.

Next, test environments must have the ability to explore failure boundaries that are not intuitively obvious and at present, we cannot explore the state space of a model in a methodical manner via an automated test capability without extrapolation. This includes the ability of test ranges to support autonomous systems' testing for boundary conditions, with appropriate safety protocols in place, which is a challenge. While non-generative AI has information on structure and boundaries that will help inform drivers and important units, generative AI based on ML poses stochastic challenges. In general, it is difficult to learn or do enough testing to learn the catastrophic and other impactful failure modes in these types of systems. T&E cannot handle testing that requires extrapolation of behavior due to potentially catastrophic results. And the results of black box models that lack visibility into how AI is making meaning are near impossible to extrapolate. This inability to decode the direct links from inputs to outputs impacts evaluation and requires more data and development.

Furthermore, we need to explore the relationship between design and the burden of testing, to comprehend the boundaries that might permit some amount of extrapolation. Participants noted that even interpolation has instability. Can we move toward extrapolation, particularly when testing must cross the range of potential behaviors of a system? Or when testing moves beyond two-dimensional learning to accommodate multiple variables. As the number of variables influencing the AI-A model increases, at some point knowledge about why a system is behaving a certain way becomes indecipherable, particularly when compared across differing environments.

This also leads to the notion that factors that matter to performance will change over time as a system learns. Such behavior can be painful and difficult to quantify or observe with certainty. Given this truth, test and evaluation of AI-A enabled systems cannot be viewed as a single evaluation at a point in time, but a critical capability that must be monitored over time as operational factors and AI capabilities evolve.   Moreover, success of an AI component may depend on the requirements of a given mission and depend upon the type of model employed and how it is integrated. This includes the needs of the operational scenario at hand and the level of accuracy required of the system. Take, for example, the differences in accuracy required to put a missile on a boat with a tight desired mean point of impact (DMPI), versus some kind of supply logistics delivery that doesn't need to meet strict space/time attributes. Mission and context are important. The impact of failure is important.

Breakout Session #3: System T&E

Breakout Objective(s) and Discussion Topics:

Objective 1: Compare and contrast System T&E versus Model T&E for AI-A systems.

Breakout session participants first completed a 'compare and contrast' of System T&E versus Model T&E based on four T&E considerations that are common to both focus areas:

- T&E Objectives. What do we aim to learn about the AI-A capability and performance during Model T&E vs System T&E? How do we ensure clear alignment of objectives to measures and requirements?
- Test Designs. What conditions can be (or should be) tested during Model T&E vs System T&E? How should verification and validation be approached? How do we account for continuous test?
- Data Characteristics. What types of data can be (or should be) used during Model T&E vs System T&E? How important is operational data? How do we ensure data will be sufficient to support performance assessment?
- Resource Intensity. How do resource needs compare for Model T&E vs System T&E, to include: data, computing infrastructure, test ranges, AI/ML skilled workforce, etc.?

Objective 2: Explore current strengths, weaknesses, opportunities, threats to plan and execute system T&E for AI-A systems.

Next, participants were asked to isolate the unique aspects of System T&E, and characterize DoD's baseline abilities to plan and execute this focus area for AI-A systems. This activity was conducted as an informal "SWOT" analysis: **S**trengths, **W**eaknesses, **O**pportunities, **T**hreats.

Key Takeaways:

Similar to other breakout sessions, ***the need for common lexicon*** around key terms that include "model" and "system" emerged as a resounding takeaway.

Additional key takeaways included:

(1) T&E for AI-A Systems should be an iterative process, and integration across the T&E phases is key.

(2) Too much attention is being paid to figuring out how to plan / execute T&E, without enough consideration of mission context and intended use. Most often, AI-A developers, program managers, and testers do not have a clear statement about the missions the AI-A will perform, the context of those missions, and relevant CONOPS

(3) Human-systems integration (HSI) is a critical part of successful development and operations of AI-A systems, but it is not currently emphasized in T&E. The T&E community should work closely with the operational community and acquisition program managers to encourage and increase use of HSI engagements across the lifecycle.

Participants also recommending adopting industry best practices, but recognizing there is no one-size-fits all, DoD needs to understand industry goals and objectives before deciding whether it will work for DoD.

Discussion Summaries:

While System T&E has been identified as one of six focus areas of the AI-A System T&E Lifecycle, it is important to realize that this should not be viewed as independent. The AI-A System T&E Lifecycle is a continuous and iterative process, and various focus areas will overlap throughout T&E planning and execution. System T&E specifically is expected to have significant overlap with Model T&E, with both focus areas playing significant roles in evaluating the maturity and performance of AI-A system sub-components prior to full integration, as well as during sustainment. Often, Model T&E and System T&E will be conducted by contractors and/or government developmental test and evaluation (DT&E) organizations, but this can vary by acquisition pathway.

Discussions during this breakout challenged participants to think carefully about where System T&E and Model T&E focus areas overlap versus how they serve unique purpose within AI-A System T&E Lifecycle. This level of understanding on the relationship between the focus areas is necessary to optimize their overall planning and integration throughout the lifecycle.

Compare and Contrast, System T&E versus Model T&E

Discussions during the "compare and contrast" of System T&E versus Model T&E revealed significant overlap between the two focus areas, while also identifying clear demarcations in how they contribute to the overall AI-A System T&E Lifecycle. Overall:

> **Model T&E** has the primary objectives of assessing the reasoning and drift of the AI-A system models. T&E designs should emphasize fast, virtual, and agile, and should include baseline capture. The data used for Model T&E should be curated, controlled, robust data that accounts

for the full operational space and has low AI model classification error. The resource intensity and cost of Model T&E is relatively low, and could leverage testing-as-a-service.

System T&E has the primary objectives of assessing interoperability and system performance, and involves verification and validation (V&V) of the model, along with the completed training process. Here, the model has been incorporated into some type of larger system so we may need to verify the model's T&E results still hold after integration and that the larger system preforms as expected and does not lead to unexpected and problematic behaviors more frequently than our risk acceptance. T&E designs should emphasis system performance in contested environment, as well as cybersecurity, integration, and API interfaces. The data used for System T&E is expected to be operationally relevant, noisy, dynamic, and variable, and may be classified if collected from operational missions or off platforms with restricted capabilities. The resource intensity and cost of System T&E is relatively high, with increased use of LVC and M&S, and considerations of user interfaces, multiple systems models, and data scarcity.

Both System T&E and Model T&E share the objectives of contributing to the definitions and refinement of AI-A system requirements and metrics. T&E designs should be mission-relevant (including rare unexpected events) and risk-based, and should emphasize testing of robustness, cybersecurity, and safety. The use of M&S and LVC is common for both focus areas, as is the use of both synthetic and real data involving some levels of variability. Factors impacting resources: domain expertise, continuous training and data analysis, digital ranges and tools, test harnesses, cloud versus on prem, and static versus data pipelines.

Systems T&E: Strengths, Weaknesses, Opportunities, and Threats

Building on the unique contributions of System T&E identified in the previous activity, the breakout participants explored DoD's abilities to carry out this focus area based on current acquisition practices, workforce, infrastructure, etc. These discussions produced an informal SWOT analysis, which is summarized below in Table 4 and Table 5.

Within the tables, the four elements of the SWOT analysis are organized as strengths (S) and opportunities (O); and weaknesses (W) and threats (T). The (*) represents the number of participants who identified the item as important.

It is important to note that the weaknesses and threats are not indicators of criticisms or failures in existing DoD acquisition. Instead, these are areas that were identified as being part of the critical path of success for System T&E of AI-A systems, and as such require further discussion by this community of interest and T&E leadership on what could be done to address them.

Table 4. Strengths (S) and Opportunities (O) for System T&E

Test Objectives	Test Design
• Common lexicon (O) ********** • Adversarial red teaming (O) *** • Translate words to metrics (O) ** • Examples of T&E past successes of non-AI through systems engineering (S) **	• Bring warfighter in early (O) ******** • Leveraging existing standards (O) *** • HSI execution (O) *** • Looking at what level we test a system (O) ** • Design pairing AI with end users and their systems (O) ** • Model test cases that can't be done live due to security (O) **
Data Characteristics	**Resource Intensity**
• Synthetic data is cheap (O) ****** • Quantity (S) ** • Data lakes (O) **	• Collaboration / workshops (O) *** • Experience from AI-A pathfinders (O) *** • Farm system (students) (O) ** • More buy-in from diverse stakeholders (O) * • Leverage best practices from industry (O) ** • Virtualization of systems (S) **

Table 5. Weaknesses (W) and Threats (T) for System T&E

Test Objectives	Test Design
• Taking too much time talking about systems as opposed to fielding them (W/T) ***** • Definition of assumptions (W) **** • Clear goals and objectives (W) **** • Data poisoning (T) ***	• Lack of requirements (W) **** • Live testing is not required since we have a model (T) **** • Sufficient or enough testing for AI (W) ** • Design limited by test infrastructure constraints (W) ** • Not considering humans as part of systems by default (T) **
Data Characteristics	**Resource Intensity**
• Data standardization (W) **** • Generate data that can capture realistic characteristics to train model (W) *** • Availability and dissemination (W) *** • Data sharing across the enterprise (W) ** • Collect quality data (W) ** • Data compromise / data negligently shared (T) **	• Difficult to hire / retain workforce (W) ***** • Lack of SMEs with domain expertise (T) ***** • Government test organizations are not set up to do model test (W) ** • Personnel experience (W) ** • Adversary experience is higher (T) ** • Classification of AI programs (SAP) (T) **

Breakout Session #4: Live Virtual Constructive (LVC) Environments

Breakout Objective(s) and Discussion Topics:

Objective: Explore the requirements for LVC to support T&E of AI enabled and autonomous systems.

Topics of discussion included:

- The relationship between LVC and OT&E
- The potential impacts of LVC environments
- The need for additional T&E range capabilities

Key Takeaways:

LVC, DT&E, and OT&E are not mutually exclusive phases of the AI-A System T&E Lifecycle. LVC should be used for DT&E and OT&E planning and augment testing. Key benefits or attributes include:

(1) LVC may be able to help overcome OT&E limitations, to include safety by helping test failure modes and conditions that would expose humans to danger.

(2) LVC, once created, may provide lower cost testing and allow for many more runs faster than OT&E.

(3) LVC offers multi-phase and multi-domain development opportunities for iterative learning and systems updates.

The potential impact of LVC environments is incredibly strong. But they require high levels of data fidelity and quantity that comes with high costs. For example, in order to link LVC environments, high speed computing capabilities will be required at a level we do not understand.

Discussion Summaries:

The relationship between LVC and OT&E

Table 6 provides a list of attributes that describe the relationship between LVC and OT&E:

Table 6. Relationship between LVC and OT&E.

LVC – OT&E Relationship	Example Attributes
LVC and OT&E are not mutually exclusive. LVC should be used for OT&E planning and augment or be part of later testing.	• LVC is preferred when identifying edge cases, though OT&E can introduce edge case variability • If LVC "lessons learned" improve, they may be able to predict OT&E outcomes • LVC instrumentation is easier than OT&E • LVC may be able to help overcome OT&E testing limitations, to include safety by helping test failure modes and conditions that would expose humans to danger • LVC, once created, may provide lower cost testing and allow for many more runs faster than OT&E
LVC offers several unique benefits on its own.	• LVC can combine the fidelity of live tests with the ability to update software in DEVSECOPS. For example, it can explore flying major weapons systems hardware alongside rapidly developing software • LVC allows modularity in constructs and can create or introduce new models and components that will more easily integrate • LVC offers multi-phase and multi-domain development opportunities for iterative learning and systems updates
The creation of LVC test environments, along with execution of LVC testing, faces several unique challenges.	• The earlier connections can be made with system operators to impact LVC design, the better • LVC validation requires a larger extrapolation between environments than OT when it comes to reflecting an operational mission. • LVC environments need to improve the incorporation of human factors to gain user-buy in and elicit realistic feedback. • LVC testing must be executed in an environment that is independent from a model's development environment. • When testing in an LVC environment, designers need to ensure that LVC data latency mirrors the experience operators will have in the field.

The potential impacts of LVC environments

While the potential impact of LVC environments is incredibly strong, they require high levels of data fidelity and quantity that comes with high costs. For example, in order to link LVC environments, high speed computing capabilities will be required at a level that requires further exploration and analysis. ~~we~~

~~do not understand~~. Data capture, to include synthetic and generative, needs to be followed and saved from past events. At the same time, LVC allows for fine-grained control of inputs and conditions, making a powerful case for traceability between inputs and outputs. That said, there are still many things we do not understand or know how to execute. Here are ***several questions and challenges that are left unanswered*** currently.

1. How do we build LVCs that are adaptive to new instrumentation and foster requirements that aren't fettered and prescriptive?
   - At present we do not enable adaptive models but rather custom built restrictive LVCs.
   - We do not have active LVC standards. Standards may change depending on risk-based scenarios.

2. Who will bear the cost of standing up and maintaining LVC environments?
   - LVC environments are not yet cloud based and instrumentation between live and virtual world needs to mature from disk handling procedures.
   - Ranges need to connect across test agencies to JITC and back haul data to closed environments for test/fix/test capabilities.

3. How will we determine the right fidelity to make virtual and constructive environments reflective of the real world?
   - There is a possibility that requirements are so stringent that the scope cannot be met and an environment will not see completion.
   - We don't know LVC boundaries and how to leverage LVC in system maturation and T&E processes to assure and field trustworthy systems.

4. How will we incorporate live agents?
   - LVC environments require greater adversarial testing.

5. How can systems play with one another across ranges and how can data be shared?
   - Interoperability of ranges will be critical for future success. The capability to look at different systems across different domains in LVC environments does not exist.
   - There is a need for an LVC ontology and taxonomy. While past data is lost, future data could be more easily curated and leveraged by organizations like universities, industry, government.

The need for additional T&E range capabilities

Ultimately, from a policy perspective, none of the national ranges are ready or willing to test full autonomous systems in a realistic way. Safety requirements and existing rules of control of test assets are pushing autonomous systems to less risk averse communities. For this reason, the T&E community overall is headed to portable distributed testing for autonomous systems, moving away from the 23 Major Range Test Facility Bases. The community needs environments not available at fixed ranges. We need portable kit and test harnesses that can go anywhere and that can cross security classifications.

Breakout Session #5: Operational Test and Evaluation (OT&E)

Breakout Objective(s) and Discussion Topics:

Objective: Explore the progression from Model T&E → System T&E → OT&E for AI-A systems.

This breakout session explored how current OT&E practices may need to evolve to support the AI-A T&E Lifecycle. Participants were asked to consider the following topics and questions:

- <u>T&E Objectives</u>. What do we aim to learn about the AI-A capability and performance during OT&E, that builds on learning from Model and System T&E? How do we ensure clear alignment of objectives to measures and requirements?
- <u>Measures and Metrics</u>. How do we evolve our current approaches for assessing and validating system performance (e.g., effectiveness, suitability, survivability) for AI-A systems? To include accounting for unique HSI and cyber considerations, as well as the need for continuous testing of model and system performance?
- <u>Test Design and Methods</u>. What conditions can be (or should be) tested during OT&E, above and beyond Model and System T&E? Will certain test methods have larger significance in the T&E of AI-A systems, compared with other acquisition pathways (e.g., sequential test, combinatorial test, Bayesian, LVC, M&S)?
- <u>Data Pipeline</u>. How do we capture the data from OT for evaluation of AI-A components? Understand operational relevance of the AI-A components? What do we need ranges to capture, versus synthetic? What is the role of OT data captured in the feedback loop for training?

Key Takeaways:

Overall, while there are several opportunities to evolve OT&E practices to support the T&E of AI-A systems, most of these opportunities are not unique, and also apply to the OT&E of complex systems. That said, the implications and risks of not evolving OT&E practices are much greater for AI-A systems.

Aspects of AI-A systems that raise concerns about current OT&E practices include:

(1) The elevated importance of conducting earlier operationally relevant testing with operators
(2) The difficulties defining and assessing requirements that evolve over time with the system
(3) The increased need for follow-on testing of system performance, into sustainment
(4) The difficulty measuring trust

The breakout participants recommended **developing a framework to differentiate metrics** as a way to help manage changes over time and mitigate risk throughout the course of the AI-A system lifecycle.

Discussion Summaries:

T&E Objectives

Discussion on T&E objectives focused on the importance of OT&E building on the earlier test data and results generated during Model T&E and System T&E. Participants agreed that in general, the need to leverage previous testing through an end-to-end and iterative T&E approach is not unique to AI-A systems. Where the distinctions come into play for OT&E of AI-A systems are:

- Building upon developmental test (DT). Making sure OT&E is informed by DT, and making sure DT data and results are trustworthy with known value propositions and limitations.
- Operator involvement and operationally relevant T&E. Including operator touchpoints in testing prior to OT&E.
- Follow-on testing. Determining how must testing is enough during OT&E to anticipate how it will perform when deployed in a new environment, then leveraging this data during follow-on testing during sustainment as the system and its environment changes.

Measures and Metrics

When discussing measures and metrics for OT&E of AI-A systems, one must account for the elevated importance of producing T&E results that are reliable, consistent, and explainable. While this may seem intuitive, several aspects of AI-A systems make this challenging, including:

- Requirements and metrics that are expected to evolve over time
- System performance that is expected to evolve over time
- Difficulties measuring abstract topics such as trust and ethics

The breakout participants recommended **developing a framework to differentiate metrics** as a way to overcome these challenges during OT&E and over the course of the AI-A system lifecycle. The framework should allow both developers and testers to determine what needs to be considered during each iteration (or sprint) in development, to include integration with other systems. It should also serve as a risk management tool to support decisions on how much risk should be underwritten on AI-A system performance during OT&E, considering not only initial system deployment but also moves to other future operating environments.

Test Design and Methods | Data Pipeline

These two topics were challenging for the breakout participants. It was difficult to explore how current OT&E practices may need to evolve for AI-A systems because similar discussions are happening on how to do better in these areas for non-AI systems (in particular, complex systems). Overall, the participants agreed that there are few differences when it comes to AI-A systems, but the differences that do exist are important and require further discussion.

Future discussions on OT&E test design and methods should include:

- Exploratory OT&E and emergent behaviors
- Integrated testing
- Human factors and trust
- Risk triage
- Edge case identification and test

Future discussions on OT&E data pipeline should include:

- Data and systems that are operationally relevant
- Results that are reproducible
- OT&E that is ongoing instead of a discrete activity

Breakout Session #6: Model Sustainment and Updates

Breakout Objective(s) and Discussion Topics:

Objective: Explore the role of T&E is once an AI-A system is fielded.

Topics of discussion included:

- Areas of responsibility for AI-A system sustainment
- Gaps in T&E during AI-A system sustainment

Key Takeaways:

Regarding sustainment and updates, T&E should have a primary role in defining considerations, metrics, and standards for capturing fielded system performance and user feedback. It also, however, needs to overhaul the concepts of initial operating capabilities (IOC) and final operating capabilities (FOC) for AI-A system fielding and sustainment.

Several additional areas of responsibility were identified that are not attributable to T&E, and largely involve retraining, defect identification, and repair.

Current practices for the T&E and sustainment of non-learning systems do not support the continuous learning of AI-A systems. Three key gaps were identified:

(1) Online learning
(2) Risk Management
(3) Requirements Specificity

Discussion Summaries:

Areas of responsibility for AI-A system sustainment

As stated above, regarding sustainment and updates, T&E should have a primary role in capturing feedback on system performance.

T&E also, however, needs to overhaul the concepts of initial operating capabilities (IOC) and final operating capabilities (FOC) for AI-A system fielding and sustainment. This includes developing tools that can cross operational environments, and conducting product improvement testing that is more agile to respond to software intensive systems.

In addition, T&E may also develop systems in which humans extrapolate from AI to see how an AI will do under different conditions when it extrapolates. This requires exploring the idea of an operator's role in collecting data in operational environments, and how that data feeds back to and becomes part of the T&E process. Participants discussed various models such as having a T&E representative deployed and collecting data during training operations or having a feedback loop from operational forces back to T&E organizations. Alternatively, service contracts could enable a tester from the developer/contracting company could conduct the data pulls and update models. Participants noted that this is not just data specific to the AI component, but includes the management of human-machine teaming and device implementation. It also depends on whether there is a service contract to maintain and support the system or whether it will be passed to a Service organization that may not have testers embedded. Regardless, we will need to account for data considerations associated with ownership, security, and its potentially proprietary nature. The ability to pull data from the field was identified as a gap that could limit the operational impact of AI. Participants noted the T&E community could assist in breaking those barriers where possible.

DT&E and OT&E have a combined responsibility to conduct T&E on system updates, as well as monitor the performance effectiveness of those systems. This includes revalidating major updates and changes through verification, validation, and accreditation.

Several additional areas of responsibility were identified that are not attributable to T&E. The questions below are intended to support further discussion amongst the community of interest and/or leadership:

- Who and how will models be monitored to identify drift?
- What/who will trigger retraining of a model?
- Who will retrain a model?
- Who will pay for re-evaluation and retesting?
- Who will identify relevant model performance training data for relooks?
- Between test events, who will conduct ad hoc testing, red team, and test for vulnerabilities?
- Who will handle emergent defects?
- Who will conduct repair or retraining?

Gaps in T&E during AI-A system sustainment

Breakout participants strongly agreed that current practices for post-sustainment T&E of non-learning systems do not support the continuous learning of AI-A systems. Non-learning systems are designed for traditional mechanical or deterministic systems that do not adapt or evolve over time. As such, ***significant gaps exist in the T&E and sustainment of learning systems, including AI-A systems.*** Three areas were identified as having significantly unique, unaddressed considerations and implications for AI-A systems.

1. Online learning
2. Risk management
3. Requirements specificity

Discussions on these three areas are summarized below, including examples that helped explore the question “how do we get there?”.

<u>Online Learning.</u> The community requires a model to predict how a system will learn something in the field. Unfortunately, such predictions cannot yet be made in a lab, let alone the field environments. Some individuals called this adaptive learning vice online or offline learning. Air Force representatives highlighted it is pursuing the end goal of online learning and in the interim pursing the processes and methods to enable daily model retraining. They described a process flow where, XQ-58’s drones collect data during a flight, land and then share that info across training aircraft before going out the next day with new data. They are currently working to send data during flights to other aircraft. In this way, the squadron limits the risk of online learning with autonomous systems, but enables rapid capability

updates. During their testing, perceptions change but most behaviors are not allowed to adapt - only certain models. Perception systems may be retrained more frequently than the behavioral/flight models.

Breakout participants then discussed online learning false alarm radars and the possibility that certain decisions can accept risk for this kind of adaptation, particularly those with limited impact and criticality. For example, if a human can easily override the new perception or if there are other redundant clarifying sensors.

Risk Management. Because there is diverse range of decisions that exist – risk management changes in the presence of online learning. Regardless, complex issues with emergent behaviors and unintended consequences and limited governability must have the ability to roll back or reset remotely. Participants noted a lack of legal guidance in this field or circumstance.
Model ownership and correction control are not standardized in contracts and may be unclear under certain circumstances. Currently, contracts are ad hoc and differences in the details may have large and problematic impacts.

One example helped to reframe the concept of risk acceptance for the team. We were asked to think about how generative AI could change the game for mine sweeping operations. Specifically, JHUAPL introduced the impact of a test system that might be able to identify a new brand or version of a mine for which no existing tactics, techniques and procedures (TTPs) exist. To a learning AI system, the unidentified object is a novel thing to the environment. As part of a standard workflow, the system could flag the object and bring it to the attention of a human operator. The operator could then in real time determine the validity or potential of the AI determination. At the same time, if the AI is correct in its new label, that information could then be used to retrain the system offline. Over time, this retraining process could grow and mature. It also provides the opportunity for the operator to reject the AI's incorrect conclusions and revert to the previous version when it learns ineffectively.

Requirements Specificity. For reasons stated above, from the T&E perspective, specificity of requirements and what is being trained need to be very specific. Participants cited the example of moving from ChatGPT 3.5 to 4.0, where the large language model (LLM) capability in specific areas (e.g., math) decreased. If many models are training on many ships, how will that information be fused? How will various operational environments and specific instances of new object classification merge? Will online aggregation or transfer learning make a difference, or will each ID need human adjudication? If systems are continuously learning, we must come up with a way to do continuous VV&A, not just periodic, but continuous. Right now, we are crawling in this area and cannot allow change to happen in the field. The run phase would enable change. But we are a long way from this.

Conclusions

Overall, the workshop succeeded at bringing the T&E community together to discuss a path forward on testing AI-A enabled systems. This workshop report provides numerous key finding and next steps. There were a few recurring themes from the breakouts that should be prioritized to move the conversation forward. These include:

- Need for common lexicon agreed upon across the AI-A and T&E community,
- Importance of operator involvement & operationally relevant testing early, and
- Need to accommodate requirements and metrics that evolve over time with the system.

While there were numerous areas where future work was needed there was broad consensus that T&E must update its processes and timelines to be responsive to AI-A enabled system development. There was also broad consensus on the important role T&E will play in getting these capabilities to operators in a way that they can be trusted.

Appendix A: Additional Resources

- T&E Preplanning: Scoping, Requirements, Acquisition; Strategies: CSET's NIST AI, RMF Profile Template
- Ms. Standard Cyber DTX Service WG (DPE&A), Have Cyber Brief for IA&A Systems
- CDAO and T&E Frameworks, CDAO Jon Elliot
    - JATIC CDAO – Jon Elliott and David Jin
- Investments – "Future Flag" tech exercises hosted by AFRL/RI and 174th ANG Sq.; POC Pete Lamonica AFRL/RI, Rome, NY
- Research, Case Studies, Smart Sensor, Test Plans, Education; JHU for CDAO POC – Jane Pinelis;
- JATIC's AI Assurance Toolbox and T&E Workflows; POC – David Jin, PM and Ari Kapusta
- MITRE ATLAS, atlas.mitre.org, ML Vulnerability list
- COGNITIVG Electronic Warfare, RF CCB sub-committee; POC – Brandon Stringfield GTRI/TETRA
- Investments – AD4x – Autonomy, Data and AI Experimentation Proving Ground; POC – Col. Tucker Hamiliton and Lt. Col. Dan Riley, AFWERX Autonomy Prime; POC – those above and Lt. Col. Bryan Ralstan
- Assurance Case Tools/Demo; POC – Dave Sparrow dsparrow@ida.org
- Invest: TRMC ADTR Toolkit; POC – Ellen Preiss
- Investment: TRMC ADAS, AAIT, OA2I; POC – Christopher.f.lynch.civ@mail.mil
- ARCEM algorithm evaluation for EW; POC – AFRL Tony Buchenroth anothony.buchenroth@us.af.mil
- Naval Autonomy Test System (NATS) by the Navy Autonomy System Test Capability Project (TRMC CTEIP)
- Investment: LVC Construction T&E, Army Test and Evaluation Command is resourcing development of multi-domain operations (MDO) LVC test architecture; POC – Robert Duffy ATEC P4
- MLOps for Defense (Mostly EW) Applications; POC – GTRI Austin.ruth@gtri.gatech.edu or Jovan Monroe
- Robust AI Test Experiment; POC – NSWC Tyler Fitzsimmons and ARMORY Model Evaluation Framework – DARPA
- T&E and Safety with Army Safety; POC - Army AFC, DEVCOM AC, Ben Schumeg
- Data Analysis and Assurance; POC - Army AFC, DEVCOM AC, Ben Schumeg
- AI Trust and Assurance; POC - Army AFC, DEVCOM AC, Ben Schumeg
- DoDD 3000.09 Analysis; POC - Army AFC, DEVCOM AC, Ben Schumeg

- Maven has a mandate to provide T&E as a service. Could maybe leverage more widely; POC Ashley Suiter, NGA
- TES for Software Pathway from Joint DEO
- JSE – Navy/AF (F35), Digital Test Training Range – AFTC, AFOTEC/TRMC, AFOTEC

Appendix B: Workshop Participants

First	Last	Organization
Nicholas	Adams	Air Force Operational Test and Evaluation Center (AFOTEC)
Kristen	Alexander	DOT&E
Matthew	Alsleben	AFWERX
Miriam	Armstrong	IDA
Shannon	Arnold	OSD Principal Director Trusted AI and Autonomy
Logan	Ausman	IDA/CDAO
Peter	Ballentine	AFLCMC/WLQ (Future Tanker Program Office)
Oliver	Barham	NGA
Chad	Bieber	CDAO Assess and Assure
Curtis	Bonham	HQ STARCOM S2/3V
Christopher	Borkowski	
Thomas	Boucher	NSWC Dahlgren
Joy	Brathwaite	Aerospace Corporation
Georgianna	Campbell	Naval Information Warfare Center Atlantic
Ryan	Caulk	USSF/STARCOM
Joseph	Chapa	
James	Cooke	DUSA-TE
Bryan	Davis	TETRA
Ross	Elder	40th Flight Test Squadron
Amanda	Elkins	OPNAV N942
Jonathan	Elliott	CDAO
Dave	Emerson	NSWCDD M- Department
Kelli	Esser	Virginia Tech National Security Institute
Christopher	Fairfax	Software Engineering Institute
Orlando	Flores	USD(R&E)/DTE&A
Heather	Frase	Center for Security and Emerging Technology (CSET)
Laura	Freeman	Virginia Tech
Beverley	Gable	DAU
Lynne	Graves	SAF/CNDI
Rachel	Haga	IDA
James	Hall	JITC
John	Haman	IDA
Elizabeth	Haro	Naval Surface Warfare Center, Dahlgren Division
Tony	Harris	NSWC Dahlgren
Jason	Hustedt	IDA
David	Jin	CDAO
Ian	Joyce	AFRL - RHWOH

Ariel	Kapusta	MITRE
George	Khoury	IDA/CDAO
Raymond	Kramer	AFRL/711 Human Performance Wing
Willis	Lacy	NSWCDD
Kathryn	Lahman	Johns Hopkins Applied Physics Laboratory
Cara	LaPointe	Johns Hopkins Institute for Assured Autonomy; Johns Hopkins Applied Physics Lab
Robert	Loibl	USSF SPOC 3 TES/MA
Ryan	Luley	Air Force Research Laboratory
Nix	Maegen	VT
Nicholas	Mastromanolis	ATEC
Michael	Mattarock II	Carnegie Mellon University / SEI
Kenny	McDowell	Army Test & Evaluation Command (ATEC)
Rebecca	Medlin	IDA
Charlie	Middleton	STAT Center of Excellence
Caleb	Miller	Lawrence Livermore National Lab
Mina	Narayanan	Center for Security and Emerging Technology
Maegen	Nix	VT-ARC
John	O'Donnell	711th Human Performance Wing
Austin	Omlie	Army Test and Evaluation Command
Daniel	Owens	US Army Test and Evaluation Command
Phil	Pace	IDA
Dale	Parsons	AFRL Sensors Directorate
Rohintan	Patel	Naval Sea Systems Command
Ron	Penninger	
Jane	Pinelis	JHU/APL
Carol	Pomales	MITRE Support to DTE&A
Ellen	Priess	JITC/JTA
Wade	Pulliam	OUSD(R&E)
Bryan	Ralston	
Jan	Rice	JHU/APL
Danny	Riley	AFOTEC Det 2
Stuart	Rodgers	IDA
Richard	Ross	Naval Surface Warfare Center, Dahlgren Division
Josh	Rountree	DAF-MIT Artificial Intelligence Accelerator
Anna	Rubinstein	NGA
Kimberly	Sablon	USDR&E
Benjamin	Schumeg	US Army - AFC - DEVCOM AC
John	Seel	NSWCDD
Annette	Skarhus	DoD DISA EIIC JITC
David	Sparrow	IDA
John	Stogoski	Carnegie Mellon University / Software Engineering Institute
Daniel	Suma	NSWCDD
David	Tate	Institute for Defense Analyses
Miles	Thompson	TRMC
Brian	Vickers	IDA
Robert	Waller	

Josh	Wallin	Center for a New American Security
Amy	Walters	NGA
Brian	Woolley	MITRE Corp
Qing	Wu	Air Force Research Lab