DoD Manual 5000.101

Operational Test and Evaluation and Live Fire Test and Evaluation of Artificial Intelligence-Enabled and Autonomous Systems

Originating Component:	Office of the Director of Operational Test and Evaluation
Effective:	December 9, 2024
Releasability:	Cleared for public release. Available on the Directives Division Website at https://www.esd.whs.mil/DD/.
Approved by:	Douglas C. Schmidt, Director of Operational Test and Evaluation

Purpose: In accordance with the authority in DoD Directives (DoDDs) 5141.02 and 3000.09, and the policy established in DoD Instruction (DoDI) 5000.98, this issuance implements policy, assigns responsibilities, and provides procedures for operational test and evaluation (OT&E) and live fire test and evaluation (LFT&E) of artificial intelligence (AI)-enabled and autonomous systems (referred to in this issuance as "DoD systems") acquired via the Defense Acquisition System or other non-standard acquisition systems.

TABLE OF CONTENTS

SECTION 1: GENERAL ISSUANCE INFORMATION

1.1. APPLICABILITY.

a. This issuance applies to:

(1) The OSD, the Military Departments, the Office of the Chairman of the Joint Chiefs of Staff and the Joint Staff, the Combatant Commands, the Office of Inspector General of the Department of Defense, the Defense Agencies, the DoD Field Activities, and all other organizational entities within the DoD (referred to collectively in this issuance as the "DoD Components").

(2) AI-enabled and autonomous DoD systems acquired via the Defense Acquisition System, pursuing any adaptive acquisition framework pathway, in accordance with DoDD 5000.01 and DoDI 5000.02.

(3) Non-standard acquisition systems (e.g., missile defense system).

(4) AI-enabled and autonomous DoD systems under special access controls, in accordance with DoDD 5205.07.

(5) DoD systems that incorporate AI capabilities, including machine learning, neural networks, and other forms of adaptive algorithms that enable autonomous decision-making, with a focus on supervised learning applications as the most mature area for OT&E and LFT&E procedures.

b. This issuance does not apply to reinforcement learning, generative AI, and other advanced types of AI and will be updated when OT&E and LFT&E procedures for these applications mature.

1.2. POLICY.

In accordance with DoDI 5000.98, the DoD plans, funds, executes, and reports on OT&E and LFT&E of AI-enabled and autonomous DoD systems to evaluate the operational effectiveness, suitability, survivability, and lethality (as applicable) of AI-enabled and autonomous DoD systems as they mature across the acquisition life cycle, including during operations and sustainment.

SECTION 2: RESPONSIBILITIES

2.1. DIRECTOR OF OPERATIONAL TEST AND EVALUATION (DOT&E).

Pursuant to Sections 139, 4171, 4172, and 4231 of Title 10, United States Code; Section 223 of Public Law 117-81; and DoDD 3000.09, the DOT&E reviews and approves exceptions and procedural deviations from this issuance for systems on the Test and Evaluation (T&E) Oversight List for OT&E and LFT&E.

2.2. UNDER SECRETARY OF DEFENSE FOR RESEARCH AND ENGINEERING (USD(R&E)).

The USD(R&E), for acquisition category ID programs under T&E oversight for developmental test and evaluation (DT&E) and assesses the adequacy of and approves DT&E strategies documented in the Test and Evaluation Master Plan (TEMP), T&E strategy, or equivalent document, referred to in this issuance as "TEMP/T&E strategy." For all other acquisition programs on DT&E oversight, advises the milestone decision authority (MDA) by conducting an independent analysis of test data, reports, modeling and simulation (M&S) results, and the adequacy of the DT&E plan in the TEMP/T&E strategy.

2.3. UNDER SECRETARY OF DEFENSE FOR ACQUISITION AND SUSTAINMENT (USD(A&S)).

The USD(A&S) enforces this issuance for DoD systems for which the USD(A&S) is the MDA.

2.4. UNDER SECRETARY OF DEFENSE FOR INTELLIGENCE AND SECURITY (USD(I&S)).

The USD(I&S) oversees intelligence support to the acquisition life cycle and advises the DOT&E concerning intelligence supportability requirements that affect OT&E and LFT&E.

2.5. DOD CHIEF INFORMATION OFFICER.

In addition to the responsibilities in Paragraph 2.7., the DoD Chief Information Officer coordinates with the DOT&E, the USD(R&E), the USD(A&S), and the USD(I&S) to synchronize the OT&E and LFT&E processes in this issuance with the:

a. DoD Cybersecurity Program in accordance with DoDI 8500.01.

b. DoD Strategic Cybersecurity Program pursuant to Section 1712 of Public Law 116-283.

2.6. CHIEF DIGITAL AND ARTIFICIAL INTELLIGENCE OFFICER.

Pursuant to Section 238 of Public Law 115-232, the Chief Digital and Artificial Intelligence Officer leads the development, coordination, and implementation of AI responsibility and assurance strategies, guidance, and policy, and provides access to enterprise infrastructure and platforms to enable joint data management. The Chief Digital and Artificial Intelligence Officer:

a. Establishes policy and issues guidance on definitions of requirements and testability for AI-enabled systems to implement and demonstrate adherence to the DoD AI Ethical Principles established in the February 21, 2020 Secretary of Defense Memorandum and the DoD Responsible AI Strategy and Implementation Pathway.

b. Issues guidance, methodologies, and best practices on T&E for AI capabilities in DoD systems.

c. Coordinates with the USD(R&E) and the DOT&E on developing and using common tools and infrastructure for T&E and verification and validation (V&V) of AI capabilities in DoD systems.

2.7. DOD COMPONENT HEADS.

The DoD Component heads follow the procedures outlined in this issuance through:

a. Component acquisition executives.

b. Program managers.

c. LFT&E organizations.

d. Their designated operational test agency (OTA) or operational test organization (referred to in this issuance as "OTA").

SECTION 3: OT&E AND LFT&E OF AI-ENABLED AND AUTONOMOUS DOD SYSTEMS OVERVIEW

3.1. SCIENCE- AND TECHNOLOGY- BASED OT&E AND LFT&E OF AI-ENABLED AND AUTONOMOUS SYSTEMS.

a. The planning, execution, analysis, and reporting of OT&E and LFT&E (e.g., cyber survivability, cognitive electronic attack) of AI-enabled and autonomous DoD systems are based on the latest advances in science and technology (e.g., differential testing, combinatorial testing, fuzz testing) to enable the:

(1) Determination of operational effectiveness, suitability, survivability, and lethality, as applicable, of AI-enabled and autonomous DoD systems across the acquisition life cycle, including during operations and sustainment, with scientific rigor.

(2) Development and implementation of risk-based level of test assessments and mission-based risk assessments (MBRAs) to inform the scope of OT&E and LFT&E and to characterize and quantify, where possible, risks to the user and to meeting OT&E and LFT&E objectives, the acquisition program, mission engineering outcome, and DoD operations throughout the DoD system life cycle.

b. AI-enabled and autonomous DoD systems may include data processing models ranging from traditional linear models and decision trees to statistical and advanced deep learning techniques. Unlike conventional systems where decision logic is manually coded, AI-enabled DoD systems derive their logic from patterns and features within training data. AI models are often complex and stochastic and commonly have high but difficult-to-anticipate dependencies on DoD systems and operating conditions. OT&E and LFT&E of autonomous or AI-enabled DoD systems will be based on advanced science and technology and will involve:

(1) Science-based test methods to evaluate the performance of AI models (e.g., accuracy, precision, recall, robustness against adversarial inputs) and determine the operational effectiveness, suitability, survivability, and lethality (as applicable) of the autonomous or AI-enabled DoD system with scientific rigor.

(2) Data management and governance plans.

(3) Robust statistical analysis to evaluate the data-intensive nature of autonomous and AI-enabled DoD systems, including the quality and suitability of the data employed throughout the life cycle for the systems' development, validation, and sustainment.

(4) Human-machine teaming science and test approaches, including quantitative cognitive and physiological assessments of human operators to inform system development.

(5) Live, virtual, constructive (LVC) technology to support and augment human-machine teaming, interoperability, and live testing of AI-enabled and autonomous DoD systems; enable

replication of specific test scenarios; and reduce the risk of bridging the AI model testing with system integration.

(6) Tools and methods to evaluate advanced AI models, system performance, and identify any unintended performance once deployed in real-world environments.

c. OT&E and LFT&E will focus on the evaluation of the AI models as integrated with other AI models, if applicable, and DoD system components within the AI-enabled or autonomous DoD system.

d. The OTA and LFT&E organizations will use advanced science and technology to maximize the use of all relevant data (e.g., contractor T&E, DT&E, integrated T&E, OT&E, LFT&E) and M&S results to determine the operational effectiveness, suitability, survivability, and lethality, as applicable, of the autonomous or AI-enabled DoD system in support of acquisition and program decisions as the AI-enabled or autonomous DoD system may mature and adapt over time.

3.2. OT&E AND LFT&E OF AI-ENABLED AND AUTONOMOUS DOD SYSTEMS ACROSS THE ACQUISITION LIFE CYCLE.

a. OT&E and LFT&E of AI-enabled and autonomous DoD systems across the acquisition life cycle must keep pace with the development cadence of the AI model and its integration with other DoD system components and must extend into operations and sustainment.

b. OT&E and LFT&E of AI-enabled and autonomous DoD systems will be integrated into the responsible artificial intelligence (RAI) life cycle including design, development, deployment, and use, as established in the RAI toolkit. Specifically, OT&E and LFT&E of an AI-enabled or autonomous DoD system across its life cycle must include:

(1) Data Management and Validation Plan.

The data management and validation plan must include:

(a) Development and implementation of a robust and well-defined data management plan to identify and characterize the authoritative datasets for AI model training, validation, and test. AI training, validation, and test datasets should be operationally representative and applicable to the evaluation of operational effectiveness, suitability, survivability, and lethality (as applicable) of the autonomous or AI-enabled system across its life cycle.

1. Training datasets are used by the developers to train the models for a specific DoD application. Such datasets may be composed of real data, synthetic data, or both. The datasets will not be used to evaluate operational performance.

2. Validation datasets, drawn from the same population as the training dataset, will be used to assess the AI model's learning performance and tune hyperparameters of the AI model.

3. Test datasets, which may be drawn from the same population as the training and validation sets, will be used to assess the capabilities and limitations of the trained model when intended learning is complete. These datasets will be shared between developers and testers to ensure that the AI model is evaluated consistently.

4. Independent test datasets will have no common instance with the training, validation, or test sets, and will not be shared with the developers. The independent test dataset will use pre-existing scenarios (e.g., defense planning scenarios), be operationally representative, and allow for the potential for spurious or erroneous inputs. The independent test dataset may contain real data, synthetic data, or both. It will consider the training and validation data (i.e., in-distribution samples) while also containing new examples (i.e., out-of-distribution samples) for demonstrating robustness to diverse and changing operational and environmental conditions.

(b) Documentation of dataset preparation, quality, governance, suitability, limitations, and a sustainability plan for data pipeline updates.

(c) Definition of the operational context, pedigree, history, and training dataset within which the AI model developers will train the AI model.

(d) Documentation of the context associated with the data and metadata that reflect the expected operational environment for OT&E and LFT&E.

(2) AI Model OT&E and LFT&E.

AI model OT&E and LFT&E may include:

(a) Functional and non-functional testing of the AI model and its learning algorithm as a standalone component to evaluate the learning process. OT&E and LFT&E planning, execution, analysis, and reporting will consider the stochastic behavior of the system due to the statistical or deep learning nature of the underlying algorithm potentially resulting in different models for the same set of training inputs over multiple training runs.

(b) Evaluation and tracking of the performance of the AI model during training and testing, including, but not limited to, security, safety, bias, variance, and instability in operationally relevant conditions, including in the presence of adversarial attacks.

(c) Mapping of AI model capabilities to operational and system requirements in support of the evaluation of AI model contributions to operational effectiveness, suitability, survivability, and lethality (as applicable).

(d) Repeats of the training phase and its V&V until the trained model reaches the required performance metrics as defined during the requirements management process. Data from the training phase may be used to inform the scope of OT&E and LFT&E.

(e) The growth of the independent test dataset to become more comprehensive over time and ensure it incorporates relevant aspects of operational realism and the operational space with statistical significance. The growth of the test dataset should be defined in requirements and test planning documentation.

(3) OT&E and LFT&E of the AI Model as Integrated with the DoD System and Its Components.

OT&E and LFT&E of the AI model as integrated with the DoD system and its components must include:

(a) Evaluation of the AI model and its interaction with other DoD system components, the human-machine team, and the DoD system as a whole. OT&E and LFT&E will vary depending on the role of the AI model and how it is integrated in operational workflows.

(b) Integrated T&E, OT&E, and LFT&E to evaluate the AI model's performance when integrated with its supporting components and sub-components.

(c) Configuration and deployment of the trained AI model to planned LVC environments, planned operational hardware and software environments, or another operational environment. Evaluation will consider potential effects of AI model compression or conversion, if applicable, when deploying AI models across different platforms or platform configurations.

(d) Verification of the AI model behavior in the operational context through the execution of operational test cases (also termed the inference model verification).

(e) Evaluation of consequences from potential system errors, deficiencies, and vulnerabilities.

(f) Characterization of emergent behavior and negative test results as the AI-enabled or autonomous DoD system learns and changes its behavior based on the input data it operates within.

3.3. OT&E.

OTAs must plan and execute OT&E in accordance with the procedures outlined in Paragraph 3.3. of DoDI 5000.98 and must:

a. Collect the data required to support the evaluation of operational effectiveness and suitability of AI-enabled or autonomous DoD systems while taking into consideration survivability and lethality effects. In addition to the metrics and measures detailed in DoD Manual (DoDM) 5000.100, operational effectiveness and suitability of autonomous or AI-enabled systems will also be informed by the evaluation of:

(1) The five AI ethical principles: responsible, equitable, traceable, reliable, and governable, in alignment with the May 26, 2021, Deputy Secretary of Defense Memorandum. See the RAI Toolkit for specifics.

(2) System safety (e.g., minimizing AI accidents, mistakes, or misuse which could cause unnecessary harm to users, equipment, or other systems), interpretability (i.e., the degree to which decisions made by the model can be interpreted by humans), and security (e.g., minimizing vulnerabilities in contested cyberspace).

(3) Human-machine integration, including, but not limited to:

(a) The tasks that the AI-enabled or autonomous DoD system will be involved in to provide a detailed understanding of what the DoD system (including the human) is intended to do, how it is intended to do it, and what might prevent it from achieving its goals.

(b) Usability, including, but not limited to, user experience, interfaces, and user-friendliness of the autonomous or AI-enabled DoD system.

(c) Accuracy to measure how often the DoD system correctly performs its assigned tasks, how often the system performed unintended tasks, response time to measure how quickly it responds to user or other relevant inputs, and workload to measure how much work the system can handle.

(d) The degree to which the user can predict and interpret actions of the autonomous or AI-enabled DoD system across the operating envelope, including when the system exits the expected performance envelope.

(e) The existence of confidence between the user and the DoD system including the ability of the user to interpret the DoD system decisions.

(f) The user's ability to recognize and react to unexpected DoD system behaviors to prevent unintended system performance and mission outcomes.

(g) The evaluation of instructions and resources provided to the users to properly operate, maintain, and support the DoD system.

(h) Operational effectiveness and suitability of human-machine teaming.

b. Define OT&E adequacy criteria, while considering the data-intensive character of AI-enabled and autonomous DoD systems. Examples include, but are not limited to:

(1) Metrics that measure test adequacy based on the diversity of the test dataset (i.e., the degree of similarity or dissimilarity between the test dataset and the training dataset).

(2) Metrics that evaluate data sufficiency based on the presence or absence of feature interactions in a dataset compared to the operational environment (i.e., alignment between a test dataset and the operational environment, including the measurement of how close the test dataset represents the actual operational environment for systems that learn from data).

c. Support the evaluation of any changes to operational effectiveness and suitability, while taking into consideration survivability and lethality effects (e.g., caused by intentional model updates or model drift) during operations and sustainment.

3.4. LFT&E.

Realistic, full spectrum survivability and lethality testing for autonomous and AI-enabled DoD systems must follow the procedures outlined in DoDI 5000.98 and DoDMs5000.96 and 5000.99 and must:

a. Include counter-AI techniques, data poisoning, interrogation, evasion, and extraction.

b. Consider the expanded attack surface with respect to the data, model, and sensing capabilities of the learning and inference components of the system.

3.5. CERTIFICATIONS.

The program manager is responsible for completing required certifications in accordance with procedures outlined in DoDI 5000.98 and DoDM 5000.96.

3.6. M&S.

a. M&S required to complement the live data in support of the evaluation of operational effectiveness, suitability, survivability, and lethality, as applicable, must follow the verification, validation, and accreditation procedures outlined in DoDI 5000.61 and DoDM 5000.102. Potential M&S includes, but is not limited to, the LVC environment that the developers, OTA, and LFT&E organizations will use throughout the life cycle of development, testing, and user training as a key enabler to the successful deployment of an AI-enabled or autonomous DoD system.

b. The OTA and LFT&E organizations, in coordination with the program manager, contractor, and the T&E working-level integrated product team (WIPT), also known as the integrated test team (ITT) (referred to in this issuance as "T&E WIPT/ITT") must:

(1) Track and document configuration management hardware and software associated with LVC environments across the acquisition life cycle.

(2) Integrate the live system and environment with virtual and constructive simulation to enable thorough exploration of the operational space, explore failure states, and gain insights into operational effectiveness, suitability, survivability, and lethality (as applicable) under degraded states safely. Conduct appropriate human systems integration planning for relevant M&S activities per DoDI 5000.95.

(3) Specifically focus the verification, validation, and accreditation of the LVC environment on the fidelity of representing the DoD system performance and environmental considerations.

3.7. T&E PROGRAM MANAGEMENT.

a. Program Manager.

The program manager must follow the responsibilities outlined in Paragraph 3.7.a. of the DoDI 5000.98 and must:

(1) Support the development and execution of OT&E and LFT&E strategies and plans to enable OT&E and LFT&E throughout operations and sustainment.

(2) Ensure the OTA and LFT&E organizations will have access to training data, associated metadata, and AI model technical specifications to support OT&E and LFT&E of AI models.

(3) Develop and implement a long-term sustainability plan for the AI-enabled or autonomous DoD system and include the feasibility of developing data pipelines to update the AI components as operational missions and contexts evolve. Develop a sustainment and maintenance strategy and plan post-deployment, including the criteria that will trigger OT&E, LFT&E, and reporting on the performance of the AI-enabled or autonomous DoD system after fielding.

(4) Establish a trusted agent in conjunction with the OTA and LFT&E organizations to support the development and maintenance of the test datasets.

(5) Establish a test, data, and maintenance strategy to mitigate the potential of the input data (encountered in the operational environment) to adversely affect the AI model behavior and ensure the system will continue to perform as expected.

(6) Adhere to DoD Chief Information Officer guidance on integrating OT&E, LFT&E, and the risk management framework into the acquisition processes at the initiation of the acquisition program. Risk management framework integration guidance can be found in the DoD Chief Information Officer Library under "Cybersecurity in the Adaptive Acquisition Framework."

b. T&E WIPT/ITT.

The T&E WIPT/ITT must follow the responsibilities outlined in Paragraph 3.7.b. of DoDI 5000.98 and must:

(1) Include experts in AI or autonomous capabilities, learning models, and related data.

(2) Support the determination of AI model performance metrics for AI-enabled or autonomous systems including AI models.

(3) Define OT&E and LFT&E requirements and resources, including hardware, software, storage, and computational resources, and developer datasets used for training, testing, and validation.

(4) Inform program requirements, requests for proposals, and acquisition contracts.

(5) Support the development of V&V of trained model performance using a combination of datasets.

(6) Record issues discovered during OT&E and LFT&E to inform future test planning, and document unexpected or emergent behavior of autonomous or AI-enabled systems under test.

c. OTA.

The OTA must follow the responsibilities outlined in Paragraph 3.7.c. of DoDI 5000.98 and must, in conjunction with the program's trusted agent, identify the test dataset to ensure that independent OT&E includes test datasets not used for training of the AI-enabled and autonomous DoD system.

d. LFT&E Organizations.

LFT&E organizations must follow the responsibilities outlined in Paragraph 3.7.d. of DoDI 5000.98, and must, in conjunction with the programs' trusted agent, identify the test dataset to ensure that independent LFT&E includes test datasets not used for training of the AI-enabled or autonomous DoD system.

3.8. DATA MANAGEMENT.

The data must be managed in accordance with the procedures outlined in Paragraph 3.8. of DoDI 5000.98.

a. The program manager, in coordination with the developer, must ensure the relevant developer datasets, trained models, and associated data rights, including associated metadata, needed to evaluate model performance, are available to the T&E WIPT/ITT.

b. The OTA and LFT&E organizations, in coordination with the program's trusted agent, must withhold and reserve the test datasets for independent OT&E and LFT&E.

c. The program manager must provide the infrastructure and necessary planning to ensure that AI-specific data requirements for facilitating operational testing are properly provisioned for the test and retraining of AI models and systems. This involves recognizing the unique needs of AI data, including:

(1) Developing documentation such as model and data cards, as necessary, to facilitate alignment of operational testing by capturing AI model characteristics, use cases, and limitations.

(2) Ensuring testing data and datasets are visible, accessible, understandable, linked, trusted, interoperable, and secure.

3.9. DOT&E OVERSIGHT.

Programs on the T&E Oversight List for OT&E and LFT&E must follow the OT&E and LFT&E artifacts review and approval procedures outlined in Paragraph 3.9. of DoDI 5000.98.

SECTION 4: PROCESS FOR OT&E AND LFT&E OF AI-ENABLED AND AUTONOMOUS DOD SYSTEMS

4.1. T&E PLANNING.

OT&E and LFT&E planning of AI-enabled or autonomous DoD systems will include AI models within DoD system and the DoD system itself. AI-enabled or autonomous DoD systems will likely include multiple models, and each model may have its own associated training and validation datasets and training processes. OT&E and LFT&E will consider each model and the combined outcomes of such models in the TEMP/T&E strategy.

a. Input to the TEMP/T&E Strategy.

In accordance with DoDD 3000.09, DoDI 5000.98, and DoDMs 5000.96, 5000.99, and 5000.100, the TEMP/T&E strategy must:

(1) Be informed by the risk-based level of test assessment and MBRA.

(2) Include data cards to provide insight into data collection, processing, usage, and security practices.

(3) Describe the process for the development and maintenance of the test dataset and independent test dataset.

(4) Include model cards to provide stakeholders with knowledge of the model's capabilities, limitations, and relevant performance metrics.

(5) Include model metrics such as accuracy, precision, and recall, and how additional metrics will be derived and tailored to unique system requirements and mission needs.

(6) Include M&S (e.g., LVC) and its V&V strategy to support AI model test, system integration testing, human systems integration, human-machine teaming, and OT&E and LFT&E of infrequent and dangerous conditions.

(7) Include the human-machine teaming evaluation strategy.

(8) Document and demonstrate alignment to the DoD's AI Ethical Principles (i.e., responsible, equitable, traceable, reliable, and governable) in the context of the operational mission. See the RAI Toolkit for specifics.

(9) Include operational effectiveness, suitability, survivability, and lethality (as applicable) metrics that will need to be re-evaluated throughout operations and sustainment due to system evolution (e.g., new sensors, new data, new data feed integration) and the rapidly changing threat environment.

(10) Include the system safety evaluation strategy to assess safety concerns within AI critical functions.

(11) Include a system monitoring plan to identify and address changes to design or environment that may require additional OT&E or LFT&E.

b. OT&E and LFT&E Plans.

The OT&E and LFT&E organizations will develop OT&E and LFT&E plans supporting the evaluation of operational effectiveness, suitability, survivability, and lethality (as applicable) of the AI-enabled or autonomous DoD system, in accordance with procedures outlined in Paragraph 4.1.b.(2) of DoDI 5000.98. OT&E and LFT&E plans must:

(1) Account for aspects of the operationally representative environment that may lead the DoD system to operate outside of its expected parameters and quantified bounds (e.g., inadvertent or malicious data poisoning, hallucinations) by identifying and prioritizing test cases within the specific performance envelope(s).

(2) Detail the training and validation datasets as related to the test datasets (e.g., variations in provenance of data, coverage of data across expected operational conditions).

(3) Clarify when, how, and how often models are trained, including if they are trained during test events. As models are trained, identify the required testing, regression, and validation processes. Identify how inputs or outputs to AI components have been modified over time through data and model versioning.

(4) Identify changing aspects of the performance of the system and how those changes relate to prior testing.

(5) Clearly define real and synthetic data included at the time of test.

(6) Define the test infrastructure and its planned use with consideration for test range plans and safety.

(7) Record any M&S including LVC environments used to support OT&E and LFT&E objectives.

(8) When applicable, satisfy the requirements of DoDD 3000.09.

c. OT&E and LFT&E Input to Acquisition Contracts.

The OT&E and LFT&E organizations will work with the program manager to inform acquisition contracts related, but not limited to, the following:

(1) Obtaining access to developer configuration management network and computing environment, training and validation datasets, models (e.g., architectures and trained parameters), and training methodologies (e.g., hyperparameter selections) with sufficient detail to evaluate developer's model performance results.

(2) Obtaining access to system safety assurance, software and hardware assurance, and mission performance assurance documentation.

(3) Obtaining access to software emulations and interface control documents.

4.2. TEST PREPARATION.

The OT&E and LFT&E organizations will conduct test readiness reviews in accordance with the procedures outlined in DoDI 5000.98 and DoDM 5000.96. Test organizations will identify when to conduct and revisit test readiness reviews based on the MBRA or other significant program perturbations.

4.3. TEST EXECUTION.

a. The OT&E and LFT&E organizations will execute OT&E and LFT&E plans in accordance with DoDI 5000.98 and approved test and M&S V&V plans throughout the development life cycle of the AI-enabled or autonomous DoD system.

b. The OT&E and LFT&E organizations must:

(1) Use test automation to the maximum extent possible to support planning and execution of functional, regression, performance, and security testing.

(2) Document unexpected outcomes that occur in the operational environment.

(3) Define the operational space, collect evidence, and develop the rationale for system performance across the full space to evaluate operational effectiveness, suitability, survivability, and lethality (as applicable).

(4) Collect the data for determining the capability of the system to use sensors or data feeds to perceive its environment, process associated logic, and engage in mission-relevant tasks.

(5) Support monitoring and governance of the fielded AI-enabled or autonomous DoD system, such as automated alert or disengage systems, and capabilities to revert to prior versions or safer modes if problematic performance occurs.

(6) Enable timely detection of data drift.

4.4. ANALYSIS AND EVALUATION.

The OT&E and LFT&E organizations must conduct data analysis and system evaluations in accordance with DoDI 5000.98 and DoDM 5000.96.

4.5. T&E REPORTING.

The OT&E and LFT&E organizations will generate and deliver OT&E and LFT&E reports in accordance with DoDI 5000.98 and DoDM 5000.96.

SECTION 5: OT&E AND LFT&E OF AI-ENABLED AND AUTONOMOUS DOD SYSTEMS FOR DIFFERENT ADAPTIVE ACQUISITION FRAMEWORK PATHWAYS

OT&E and LFT&E of AI-enabled or autonomous DoD systems will be included in integrated T&E, OT&E, and LFT&E events outlined for each of the adaptive acquisition framework pathways in DoDI 5000.98 and DoDM 5000.96. The scope of integrated T&E, OT&E, LFT&E, and M&S events in support of each of the acquisition decisions will be detailed in the TEMP/T&E strategy, including its integrated decision support key.

GLOSSARY

G.1. ACRONYMS.

ACRONYM	MEANING
AI	artificial intelligence
DoDD	DoD directive
DoDI	DoD instruction
DoDM	DoD manual
DOT&E	Director of Operational Test and Evaluation
DT&E	developmental test and evaluation
ITT	integrated test team
LFT&E	live fire test and evaluation
LVC	live, virtual, constructive
M&S	modeling and simulation
MDA	milestone decision authority
MBRA	mission-based risk assessment
OTA	operational test agency
OT&E	operational test and evaluation
RAI	responsible artificial intelligence
T&E	test and evaluation
TEMP	test and evaluation master plan
USD(A&S)	Under Secretary of Defense for Acquisition and Sustainment
USD(I&S)	Under Secretary of Defense for Intelligence and Security
USD(R&E)	Under Secretary of Defense for Research and Engineering
V&V	verification and validation
WIPT	working-level integrated product team

G.2. DEFINITIONS.

TERM	DEFINITION
accuracy	Closeness of computations or estimates to the exact or true values that statistics were intended to measure. For an AI-enabled system, it is measured as the percentage of correct predictions or classifications made by the model over a specific dataset.
AI	A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.
AI-enabled system	Any data system, software, hardware, application, tool, or utility that operates, in whole or in part, using AI.
AI model	A component that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.
algorithm	Mathematical process for computation based on a set of rules that, if followed, will give a prescribed result. When run on data, it creates a machine learning or AI model.
autonomous system	Systems that have the authority to make decisions independently without human intervention. Typically uncrewed, they may contain sensors, AI models, robotics, and analytic capabilities that enable an operational mode for the system to act independently.
combinatorial testing	Technique that aims to systematically test all possible interactions of input parameters of the system under test enabling testers to perform pseudo-exhaustive testing.
data card	A document for a dataset that provides explanations of processes and rationale that shape the data and consequently the models. Examples include upstream sources, data collection, and annotation methods; training and evaluation methods; intended use; and decisions affecting model performance.
data poisoning	Type of attack that involves manipulating AI data used in training, validation, or testing. The goal of data poisoning is to produce undesirable outcomes.

TERM	DEFINITION
differential testing	Compares the outputs of similar implementations and treats the inconsistency among implementations as indications of faulty behavior.
emergent behavior	Unexpected patterns, actions, or behaviors that arise from the interactions of simple components within a complex AI system. These behaviors are not explicitly programmed but emerge from the system's dynamics.
equitable	AI capabilities that are fair and operate in an unbiased manner.
evasion	A deliberate attempt to manipulate or deceive AI systems to circumvent their intended functionality or to produce inaccurate results. Can be used for various purposes, including bypassing security mechanisms, fooling automated detection systems, or undermining the reliability and trustworthiness of AI-powered technologies.
extraction	The process of retrieving specific information or features from data using AI techniques.
functional testing	Testing that verifies the system's expected behavior given a set of inputs or actions. These tests (sometimes referred to as "system tests") verify delivery of user value by exercising the entire system.
fuzz testing	Identifies potential vulnerabilities and assess the robustness of a software system by generating random and unexpected inputs for the system under test.
governable	AI capabilities that fulfill their intended functions while possessing the ability to detect and avoid unintended consequences and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.
hallucination	Generated content that is nonsensical or unfaithful to the provided source content.
hyperparameter	Parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. The prefix 'hyper' suggests that they are 'top-level' parameters that control the learning process and the model parameters that result from it.

TERM	DEFINITION
independent testing dataset	Data used to evaluate the AI model's operational performance and assess how well it generalizes to new, unseen data to help measure the model's ability to make accurate predictions on data it has not encountered during training. This data are not shared between developer and tester.
integrated decision support key	Defined as "IDSK" in DoDI 5000.98.
integrated T&E	Defined in DoDI 5000.98.
interrogation	Questioning or querying an AI system to obtain information, insights, or responses.
negative testing	Tests that intentionally subject a software system to invalid or unexpected inputs to assess its behavior under adverse conditions. These tests uncover vulnerabilities and weaknesses in error handling, data validation, and other critical aspects of the system's functionality. Examples include error path testing, failure testing, or fault tolerance testing.
non-functional testing	Testing to ensure the system meets quality characteristics that functional testing does not capture. Examples include performance, security, and usability tests.
model card	Short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions that are relevant to the intended application domains. Discloses the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information.
precision	Measurement of the correctness of the positive classifications made by the AI model. This can be calculated by dividing correct positive predictions by the total number of positive predictions (both true and false positives).
recall	Measurement of how often the AI model correctly identifies positive instances from all actual positive samples in the dataset. This can be calculated by dividing the number of correctly identified cases by the total of true positives and false negatives (missed cases).

TERM	DEFINITION
reliable	AI capabilities that have explicit, well-defined uses, and the safety, security and effectiveness of such AI capabilities will be subject to testing and assurance within those defined uses across their entire life cycles. AI capabilities that generate consistent and accurate outputs, can withstand errors and recover quickly from unexpected interruptions of misuse.
responsible	AI capabilities and DoD personnel that exercise appropriate levels of judgment and care in a responsible manner.
RAI life cycle	Tools, policies, processes, systems, and guidance which synchronize enterprise RAI implementation for an AI product throughout the acquisition lifecycle via a system's engineering and risk management approach.
RAI toolkit	A centralized process managed by the CDAO that identifies, tracks, and improves alignment of AI projects to RAI best practices and the DoD AI Ethical Principles. Provides modular assessments, tools, and artifacts for use throughout the AI product life cycle and enables traceability and assurance of responsible AI practice, development, and use.
security	Resistance to intentional, unauthorized act(s) designed to cause harm or damage to a system and the degree to which a product or system protects information and data so that persons or other products or systems have the degree of data access appropriate to their types and levels of authorization.
stochastic	Having a random probability distribution or pattern that may be analyzed statistically but may not be predicted precisely.
test dataset	Data used to evaluate the AI model's performance and assess how well it generalizes to new, unseen data to help measure the model's ability to make accurate predictions on data it has not encountered during training.
traceable	AI capabilities that are developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedures and documentation.

TERM	DEFINITION
training dataset	Data that an AI model uses to learn patterns, relationships, and associations between features (input variables) and the target variables (the variables the model is trying to predict).
validation dataset	Data used during AI model development and tuning to make decisions about hyperparameters, model architecture, and feature selection to ensure that the model generalizes well and is not overfitting or underfitting.

REFERENCES

Chief Digital and Artificial Intelligence Office, "RAI Toolkit," https://rai.tradewindai.com/

Deputy Secretary of Defense Memorandum, "Implementing Responsible Artificial Intelligence in the Department of Defense," May 26, 2021

"DoD Adopts Ethical Principles for Artificial Intelligence," February 24, 2020

DoD Chief Information Officer, "Cybersecurity in the Adaptive Acquisition Framework," https://dodcio.defense.gov/Library/

DoD Directive 3000.09, "Autonomy in Weapon Systems," January 25, 2023

DoD Directive 5000.01, "The Defense Acquisition System," September 9, 2020, as amended

DoD Directive 5141.02, "Director of Operational Test and Evaluation (DOT&E)," February 2, 2009

DoD Directive 5205.07 "Special Access Program (SAP) Policy," September 12, 2024

DoD Instruction 5000.02, "Operation of the Adaptive Acquisition Framework," January 23, 2020, as amended

DoD Instruction 5000.61, "DoD Modeling and Simulation Verification, Validation, and Accreditation," September 17, 2024

DoD Instruction 5000.95, "Human System Integration in Defense Acquisition," April 1, 2022

DoD Instruction 5000.98, "Operational Test and Evaluation and Live Fire Test and Evaluation," December 9, 2024

DoD Instruction 8500.01, "Cybersecurity," March 14, 2014, as amended

DoD Manual 5000.96, "Operational and Live Fire Test and Evaluation of Software," December 9, 2024

DoD Manual 5000.99, "Realistic Full Spectrum Survivability and Lethality Testing," December 9, 2024

DoD Manual 5000.100, "Test and Evaluation Master Plans and Test and Evaluation Strategies," December 9, 2024

DoD Manual 5000.102, "Modeling and Simulation Verification, Validation, and Accreditation for Operational Test and Evaluation and Live Fire Test and Evaluation," December 9, 2024

DoD Responsible AI Working Council, "U.S. Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway," June 2022

Public Law 115-232, Section 238, "John S. McCain National Defense Authorization Act for Fiscal Year 2019," August 13, 2018, as amended

Public Law 117-81, Section 223, "National Defense Authorization Act for Fiscal Year 2022," December 27, 2021

Public Law 116-283, Section 1712, "William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021," January 1, 2021

Secretary of Defense Memorandum, "Artificial Intelligence Ethical Principles for the Department of Defense," February 21, 2020

United States Code, Title 10