INHR

Center for a
New American
Security

宏理 国际

MILITARY ARTIFICIAL INTELLIGENCE TEST AND EVALUATION MODEL PRACTICES

December 2024

Edited by R.S. Panwar, Li Qiang and John N.T. Shanahan

Table of Contents

Background

This set of Military AI Test and Evaluation Model Practices is the product of a two-year consultation process among Chinese, American, and international experts convened in person and online by INHR and the Center for a New American Security. Workshops to elaborate the Model Practices were hosted virtually and in Europe, Asia and North America thanks to the generosity of Founders Pledge, Carnegie Corporation of New York, and the co-sponsorship of the Royal Danish Defense College.

The goal of the consultation process was to determine whether experts from the three delegations might share consensus on certain principles and practices for test and evaluation of weapons and related military systems with significant AI components, to make those systems operate more safely, securely and responsibly. Participants in the dialogue included academics and former officials with military, diplomatic, intelligence, computer science, corporate, and legal backgrounds from the United States, China and an international delegation from Europe, Asia, and elsewhere.

Purpose

The integration of AI technologies into military systems is gaining momentum worldwide. While in most cases adoption is progressing slowly, it is evident that the deployment of AI-enabled military systems will accelerate globally in the coming years. Drawing from experiences in the private sector, we anticipate that all military forces will encounter many challenges when incorporating AI at scale.

To advance the responsible use of AI-enabled military systems, AI Test & Evaluation (T&E) – and other essential components related to the safe, lawful, and ethical employment of AI-enabled military systems – is imperative. AI T&E covers within its ambit Validation and Verification processes as well, captured in the acronym TEVV. It is essential for the international community to establish a consensus on AI T&E principles and best practices, both to promote adherence to international humanitarian law (IHL) and to reduce the global risks associated with unanticipated or unexpected failures of AI-enabled military systems.

AI introduces substantial challenges to the traditional methods of weapon system development, testing, and deployment in military settings. While some commonalities exist between the T&E processes for conventional military hardware and software systems and those for AI-enabled systems – particularly in terms of systems engineering principles – AI necessitates significant changes to T&E methodologies. The unique characteristics of AI-enabled military systems call for tailored approaches that diverge from established T&E practices to ensure comprehensive evaluation and validation. The complexity of AI T&E is intensified by the prospect of military forces adopting hybrid architectures consisting of legacy non-AI systems, new non-AI systems, legacy systems retrofitted with AI, and new weapon systems designed with AI from the outset. These systems may operate simultaneously, necessitating the consideration of cascading effects and potential emergent behaviours when multiple AI-enabled systems interact across weapon systems, command and control architectures, and cyber networks.

There is a need for states to institute explicit measures for continuously improving the safety, reliability, and controllability of artificial intelligence technology in terms of technical security and research and development operations, enhancing the ability to evaluate and manage the safety of artificial intelligence technology, and ensuring that a human is always responsible for the use of force. It is imperative that states strengthen self-restraint in artificial intelligence research and development activities and implement necessary human-machine interaction throughout the entire lifecycle of weapons based on comprehensive consideration of the combat environment and weapon characteristics. States must adhere to the principle that humans are the ultimate responsible party, establish an accountability mechanism for artificial intelligence, and provide necessary training for operators.

A goal of these model practices is to help all nations adopt a sufficient level of T&E throughout the entire lifecycle of AI-enabled systems, with the objective of promoting delivery of effective, suitable, reliable, predictable, sustainable, secure, safe, trustworthy, and resilient capabilities, in conformance with international law.

Until states gain more experience in developing, testing, and fielding AI-enabled military systems, they should be biased towards a more cautious approach favouring additional testing before any AI military technology is fielded. They should be guided by the precautionary principle: introduction of a new product or process whose ultimate effects are disputed or unknown should be avoided.

AI Testing and Evaluation Characteristics

AI-enabled systems, both civilian and military, possess distinctive features that significantly influence the T&E process. These characteristics affect not only the evaluation of AI models themselves, but also the broader systems into which these models are integrated. The unique attributes of AI require modifying traditional T&E approaches to ensure comprehensive assessment and validation. Key features that shape AI T&E include:

- Continuous Testing and Monitoring. AI-enabled systems require ongoing evaluation throughout their entire lifecycle, from initial design through long-term sustainment. This continuous approach calls for a more integrated collaboration among designers, developers, testers, and end-users. In the realm of AI-enabled systems, the concept of 'completed testing' becomes obsolete. The dynamic nature of AI models, coupled with their ability to learn and adapt, necessitates a persistent evaluation framework to ensure continued performance and reliability.

- Post-Deployment Evolution and Unpredictability. The potential for continued learning and post-deployment transformation in AI systems, coupled with their inherent opacity, introduces an element of operational unpredictability. Therefore, there is a need to evaluate probabilistic or statistically predictable (non-deterministic) behaviors, and institute processes to identify and mitigate unexpected and unanticipated failure modes.

- Dynamic Learning and Rapid Updates. AI and machine learning/deep learning systems possess the unique ability to learn directly from data without additional coding, enabling frequent system updates and, in the case of online learning, real-time adaptations. This capability requires T&E processes that accommodate continuous integration/continuous delivery (or deployment) for deployed AI-enabled systems. Furthermore, it underscores the importance of incorporating robust instrumentation in fielded AI systems to monitor and evaluate their evolving performance over time.

- Agile Governance. AI-enabled systems require a shift from traditional linear and sequential software development methodologies to more flexible and responsive approaches. Agile development methodologies and adaptive T&E principles are essential to accommodate the dynamic nature of AI systems. This iterative approach allows for continuous refinement and improvement based on ongoing testing results and evolving requirements.

- Adversarial Resilience. In conjunction with independent ‘red teaming’ exercises, it is necessary for T&E processes to incorporate specific tests for evaluating the effects and risks of dedicated adversarial attacks against AI datasets and models. This approach advances the robustness and resilience of AI-enabled systems in adversarial environments, a particularly critical consideration for military applications.

- Data-Centric Focus and Computational Demands. The foundation of AI systems lies in their data and the infrastructure required to process it. This centrality of data introduces unique challenges, including the potential for skewed, corrupted, or incomplete datasets, which can significantly affect system performance and reliability. Additionally, AI systems typically require high-performing computing infrastructure to handle complex algorithms and vast amounts of data, The data-driven nature of AI also contributes to its ‘black box’ characteristic, where the internal decision-making processes can be opaque even to developers. This feature presents significant challenges in achieving explainability and auditability in AI systems, crucial factors for military applications where transparency and accountability are paramount.

Unique Aspects of Military AI T&E

There are vulnerabilities and risks of AI-powered systems which are particularly relevant in military settings, which military AI T&E processes must address. These are as under:

- Data Scarcity in Operational Environments. The acquisition of high-quality, representative data – essential for AI/ML-powered systems – presents significant challenges in the harsh and dynamic environments typical of military operations. This scarcity introduces multiple risks, including difficulties in generating accurate training datasets, creating operationally representative test environments, and avoiding bias when AI techniques and simulation are used to produce training datasets.

- Suitability for Diverse Operational Theaters. Military systems are often frequently redeployed across varied environments. This requires a comprehensive framework for retraining and re-evaluating system performance prior to each redeployment to ensure optimal functionality and reliability in diverse operational contexts.

- Adverse Impact of Outliers. Outliers in AI-enabled weapon systems, resulting from their inherent characteristics such as opacity and brittleness, could have very adverse effects in operational environments, since human lives are at stake. T&E processes need to incorporate stringent benchmarks for minimizing the occurrence of outliers/ edge cases.

- Accelerated Decision-Making Cycles. The integration of AI and autonomy significantly accelerates the observe-orient-decide-act (OODA) Loop in military decision-making processes. This acceleration poses challenges in maintaining effective human control and increases the risk of automation bias. These factors need to be considered while designing and evaluating systems to facilitate controlled system operation.

- Risks of Online Learning. Systems capable of online learning – adapting while in operation – present unique challenges. These systems may potentially operate outside their initial T&E parameters, leading to unintended and unpredictable effects. Safeguards and monitoring mechanisms are therefore necessary to mitigate these risks, especially for systems with lethal capabilities.

- Ethical and Legal Considerations. All military systems, including autonomous weapon systems (AWS), must be vigorously evaluated for adherence to principles of international humanitarian law. Furthermore, as applicable, T&E processes must incorporate legal reviews as mandated by Article 36 of Additional Protocol I to the 1949 Geneva Conventions.

Military AI T&E Model Practices: Objective and Structure

The overarching objective of these AI T&E model practices is to foster a global dialogue on this critical topic and facilitate international consensus on AI T&E best practices and related confidence-building measures (CBMs). This approach aims to balance the need for international cooperation with the understanding that military forces will maintain discretion regarding sensitive aspects of their AI T&E processes and procedures. The goal is to foster a collaborative environment that enhances global AI safety and governance while respecting national security considerations.

The practices that follow have been formulated to address the unique challenges and considerations inherent in AI-enabled military systems. These practices are structured into three key segments, reflecting the lifecycle of AI implementation in military contexts: first, the design and development stage; second, the deployment stage of AI-enabled weapon systems; and third, considerations related to AI transparency, trustworthiness, and confidence building between states.

The practices under the design & development segment relate to issues which impact T&E processes such as data integrity, data provenance and difficulties of data collection, and hence the need for synthetic data; transparency and interpretability especially as related to critical decision support systems (DSS) and AWS; constant integration between developers and military practitioners for optimum performance in a complex environment; continual T&E triggered by frequent redeployment in a variety of operational settings; resilient handling of edge cases especially for high-risk systems; importance of modelling and simulation to cater for data starved use cases; the need for ensuring informed oversight by commanders and civilian leadership; the imperative of human-system integration particularly in the case of AWS and critical DSS; the impact of large language models (LLMs); and the requirement of compliance with international law particularly IHL.

Practices in the deployment segment deal with issues emerging in a complex networked environment; unanticipated and catastrophic errors which might occur in harsh and dynamic operational scenarios, particularly in relation to strategic C2 systems; the need for continuous monitoring and remediation especially in systems incorporating the online learning feature; and the importance of data collection and management throughout a systems' life cycle.

The final segment on standards, incidents and confidence-building includes practices which highlight the need for drawing on existing professional standards pertaining to military and civilian contexts; the desired role of the UN and other multi-lateral organisations; sharing data and processes for dealing with incidents; training programs on risk-mitigation techniques; sharing T&E standards, policy and doctrine; suggesting proposals for a legally or politically binding instrument on T&E; and the imperative of adhering to the precautionary principle.

Glossary of Military AI T&E Terms

The model practices have been framed in concise language. In so doing, certain AI and military related terms have been used which might warrant explanation. While definitional rigour for these terms is desirable, for the purpose of this document it is considered sufficient if the implications of these terms are given out with clarity. A glossary of such terms along with their intended meanings is attached as an appendix to this document.

The 22 Military AI T&E Model Practices

I	Design and Development	
	A	AI T&E must include test, evaluation and assessment data obtained under conditions as close as possible to the conditions expected during operational deployment of the system, ideally based on real-world data. Data may be limited due to lack of collection opportunities against military targets or by adversarial actions to prevent collection, so operational data may have to be supplemented with synthetic data. Measures should be taken to assure custody, provenance, and quality of training, testing, and validation data used.
	B	Choices of test methods used should be informed by the extent to which algorithms and components of an AI system are interpretable and understandable and that these can be assessed through a robust T&E process with clear performance indicators and evaluation metrics. This is especially relevant for critical military decision support systems and lethal autonomous weapon systems (LAWS).
	C	The design and development process for AI systems should incorporate T&E requirements from the beginning. AI T&E must account for the differences between traditional software and AI, as well as for algorithmic and operational testing. As operational environments are more complex, uncertain, and less widely understood, militaries should ensure T&E plans account for the entire AI lifecycle, to include sustainment. AI T&E requires continual integration among developers, testers, and military practitioners to achieve predictable and reliable test outcomes.
	D	T&E of AI-enabled military systems should be viewed as a continual process. It should occur before and after a system is deployed, as well as before every re-deployment to a different operational environment, until a system's retirement. While T&E for periodic AI model updates and updates necessitated on operational re-deployment or changing operational conditions may differ in depth and breadth from T&E prior to initial deployment of an AI-enabled system, design plans must account for continual T&E (as part of continuous integration/continuous deployment or delivery).
	E	AI T&E should include testing under real-world conditions with due regard for the resilience and robustness of the system to include appropriate handling of edge cases and boundary conditions in harsh, uncertain, and dynamic operational environments, error correction and identification, and rollback/failsafe modes especially in high-risk LAWS.

The 22 Military AI T&E Model Practices

	F	T&E plans should specify when and to what extent modelling and simulation will be used to test AI systems and how this kind of testing will be validated, especially for systems designed to function in environments in which adversaries are expected to deny or deceive fielded AI models. For each system, it is necessary to be clear at what level of fidelity the behaviours of such system need to be tested through modelling and simulation, to include use of "digital twins."
	G	T&E plans should include how testing results and system performance will be communicated to all relevant stakeholders. T&E processes should support informed oversight by appropriate military commanders responsible for their deployment and civilian leadership.
	H	Human-system integration and/or human-machine teaming should be considered as an integral component of T&E design. A key aspect of T&E for military AI systems is determining the ability of human operators to supervise AI systems under operational conditions, and to what extent, to ensure that the observe-orient-decide-act (OODA) loop associated with weapon systems or critical decision support systems is under human control and targeting decisions remain the responsibility of human commanders and operators.
	I	Given the potential integration of large language models (LLM)/generative AI into military systems, special attention should be paid in T&E to factors arising from LLMs, including reinforcement learning with human feedback (RLHF), fine tuning, and retrieval augmented generation (RAG), along with known LLM limitations. Military systems that include generative AI must be specifically evaluated to ensure that error modes in critical decision support and weapon systems are confined within negligible limits, which must be clearly specified and assessed during testing.
	J	T&E requirements should be designed to ensure that it is possible to evaluate system compliance with relevant legal requirements, including the obligation to conduct legal reviews of a system's ability to be used in compliance with international humanitarian law and other relevant international law instruments.

The 22 Military AI T&E Model Practices

II	Deployment	
	K	T&E should assess not only the performance of components and subsystems of an AI-enabled military weapon or decision support system, but also overall AI system performance and the integration of these components, subsystems, and any external or pre-existing platforms. This should include integration or combination of new systems and updates with previous components, platforms, or systems, especially in a networked environment.
	L	T&E plans and systems documentation should identify a rigorous process by which, prior to deployment of a military AI system to a new operational context or when there are significant changes in the operational environment, hazards are identified, analysed, and remediated. Plans should establish the conditions and processes for updating fielded models, to include tactical unit roles and responsibilities.
	M	T&E plans should identify high risk catastrophic errors that could occur during operations and how these may be prevented, detected, and remediated, especially in the case of strategic command and control systems. It should also identify how unanticipated errors will be handled. T&E plans should deploy "red teams" to challenge assumptions and otherwise attack the underlying logic of the system's design and deployment to identify unanticipated errors and remediate them before deployment.
	N	T&E plans should establish how to evaluate if deployed military AI systems continue to meet their performance goals. Appropriate corrective actions should be taken if systems do not meet these goals. Particular consideration should be given to systems that continue to learn while deployed. While online learning offers the advantage of continuous performance improvement, it introduces the risk of operating the system in untested states and increases the risks of cyberattacks by malicious actors. T&E plans must ensure that special care is taken to minimise the potential negative consequences of online learning, especially in high-risk lethal weapon systems.
	O	Collection, management, assessment, and use of data throughout a military AI system's lifecycle, including during operational deployment, is critical. Attention must be given to how data and metadata are managed during collection and to ensure that data is fit for purpose in design and as updated with collection during deployment. T&E plans should particularly consider the impact of data on AI functionality and AI safety-significant functions, with the goal of system improvement and strengthening our understanding of system reliability.

The 22 Military AI T&E Model Practices

III	Standards, Incidents and Confidence-Building	
	P	As governments work to develop and strengthen their AI T&E practices for AI-enabled military systems, they should coordinate their efforts with civilian standards, tools, and documentation and draw on professional standards from military and civilian contexts, including ISO/IEC, IEEE and other standard setting organizations.
	Q	Governments should consider what role is appropriate for the United Nations or expert-level multilateral organizations with respect to standard setting or regulating military AI with respect to technical and/or governance issues that affect T&E.
	R	Governments should engage in dialogue to learn from each other and share lessons learned from development and deployment of military AI systems, including about T&E standards, T&E's role in mitigating risks and/or "incidents," and other transparency and confidence-building measures.Governments should also consider establishing training programs on the importance of T&E in the military AI system lifecycle and to include technical, military, political and legal experts in those training programs.
	S	As part of continuous T&E over a system's lifecycle, governments and international agencies should consider establishing standards for investigation and remediation of "high consequence incidents" that occur from the use of military AI during exercises or operational deployments. These standards may include (1) the type and severity of "incidents" that should result in investigation and whether investigation should occur within or beyond national jurisdiction; (2) investigation procedures, including access to and subsequent publication or protection of classified or sensitive information about the system suspected of causing the incident; (3) mitigation and remediation procedures related to the incident; and (4) the level of transparency or disclosure that may be appropriate regarding the incident, its investigation, and mitigation or remediation procedures, taking into account the need to protect classified or sensitive information.

The 22 Military AI T&E Model Practices

	T	To promote transparency, mutual understanding, and consistent best practice, states should publicly release aspects of their processes and approaches to T&E of AI-enabled military systems. Where possible given national security considerations and to build trust and confidence, states should consider publicly releasing documentation standards, policy and operational guidelines for AI-enabled military system design; criteria used to determine testing rigor and mitigation for safety critical components; criteria used to determine severity of potential AI system accidents; legal review standards and procedures and processes to integrate AI risk into overall consideration of system and system-of-systems risks.
	U	As governments adopt military AI T&E best practices, they should consider which practices could form the basis of a legally or politically binding instrument and what, if any, might be appropriate enforcement mechanisms.
	V	Until states gain more experience in developing, testing and fielding AI-enabled military systems, they should be guided by the precautionary principle: the idea that introducing a new product or process whose ultimate effects are disputed or unknown should be approached using caution, pause, and review.

Military AI T&E Model Practices: Glossary

Test & Evaluation, Validation and Verification (TEVV)

Testing. In the context of TEVV, testing refers to the execution of specific test cases to collect data about a system's behaviour, functionality, and performance. Testing is applied throughout the life cycle, from individual units to the entire system, to verify that components function as per specifications and to validate that the system meets its intended purpose.

Evaluation. Evaluation includes analyzing and interpreting the results of testing to determine if the system meets requirements and is ready for deployment. Evaluation covers both Verification and Validation, as briefly explained below:

- Verification (Are we building the system right?): Ensuring the system meets design specifications through tests at various stages.

- Validation (Are we building the right system?): Ensuring the system meets user needs and performs effectively in its intended environment.

Not all provisions which have been included by a user in the specifications of a system might be testable in a precise manner. For instance, a user might state that a system should have simple or sleek aesthetics or a user-friendly interface. Assessment of such intangible requirements require human judgement and may not be verifiable through the conduct of tests. Thus, compliance of such requirements may not fall under the heads of Verification and/ or Validation but would still fall under the ambit of the broader term Evaluation.

Military Decision Support Systems

In this document, the term military decision support systems (military DSS) has been used to refer to all military systems which are not explicitly included as part of a weapon system. In other words, in this document the entire spectrum of AI-enabled military systems is divided into two categories: weapon systems and DSS.

Military AI T&E Model Practices: Glossary

The document also uses the term critical DSS, which is a subset of DSS. The qualifier "critical" here mostly implies its dictionary meaning. Broadly, DSS which result in the deployment of military force (positioning of forces, options for attack, etc) without directly leading to release of weapons would classify as critical DSS. Non-critical military DSS would cover applications in the area of logistics, maintenance, medical, human resource management and other such systems in the military domain.

Deployment/Operational Deployment/
Operational Environment/Conditions/Context

The terms deployment and operational deployment are largely synonymous and refer to when an AI-enabled military system is transferred to military users in operational organizations. They represent when a system moves from development and testing to operational use (operational use and operational deployment/employment are the same thing). It signifies that the operational organization has accepted the system from the organization that developed and tested the system.

The term operational environment refers to any setting in which a system is being used operationally, rather than for development, testing, exercises, or experiments. It can refer either to physical domains, such as air, land, sea, and space, or to virtual domains such as cyberspace and the electromagnetic spectrum. When an operational organization has accepted a system for use, that system is then used in the operational environment. When referring to the operational environment, there is no distinction between peacetime, crisis, or conflict. Hence, it is also frequently referred to as the real-world environment.

Military AI T&E Model Practices: Glossary

The term condition or conditions refers to the totality of the circumstances affecting how an AI-enabled military system is developed, tested, and used operationally. It is related closely to the term operational environment. It includes various factors such as weather, day or night, terrain, geographical location, whether combat operations or ongoing or not, passive or active countermeasures used to deny or defeat the system, and so on.

The term context is synonymous with operational environment and conditions. It is frequently also called operational context or the operational setting (in the AI Military Model document, when used in Section III, Tab B, the words “settings” or “purposes” would mean the same thing).

Edge Cases/ Boundary Conditions

Edge cases are scenarios that occur at the extreme ends or boundaries of normal operating conditions in a system. Edge cases are critical to test because they can uncover vulnerabilities that wouldn’t appear under typical usage and addressing them can improve robustness and reliability. They are unexpected or rare situations and can reveal hidden flaws or limitations in an AI model. For example, if you do not have pandas in your training and testing imagery data, and a panda actually appears when the system is fielded, it might yield unexpected results. If you expect pandas in the operational environment, even with a low probability, the model should be trained with panda images before the model is fielded.

Boundary conditions, on the other hand, are a subset of edge cases that specifically refer to the points at or near the edges of an allowable input range. Boundary testing focuses on verifying the correct behaviour of a system at these limits, such as just below, at, or just above the maximum and minimum values of an input or output range. They are more predictable and are important for testing the robustness of AI models. For example, when developing an AI model, one should test the image classification model with the smallest and largest possible image sizes, based on the expected operational environment.

Military AI T&E Model Practices: Glossary

The difference between edge cases and boundary conditions is as follows:

- Scope: Edge cases cover any outlier or extreme scenario, while boundary conditions are specifically about the limits of valid input ranges.

- Testing Focus: Boundary condition testing is about precise input values near the allowable range, while edge case testing may involve a broader set of unexpected or rare situations.

AI enabled systems are also subjected to testing in a similar manner. However, in weapon systems and critical DSS, the edge cases need to be kept to the barest minimum (ie, very stringent performance levels need to be specified) since these could lead to effects which are catastrophic or of high consequence, even though their occurrence may be rare.

Human Control

Human control refers to the role that humans play during the system's lifecycle, from design and development through deployment and sustainment after deployment.

For deployed systems, human control includes (a) the ability of humans to monitor and comprehend information about the operational environment and system status in real-time; (b) the degree and methods of human intervention possible at different points during system operations, including but not limited to mission planning and weapons employment; (c) the capacity of humans to predict and understand the system's behaviour and potential outcomes of its actions; (d) the assignment of responsibility and accountability for the system's actions to specific human operators or commanders; and (e) the implementation of safeguards and fail-safe mechanisms that allow for human takeover or system shutdown when necessary.

Military AI T&E Model Practices: Glossary

Networked Environment

A networked environment encompasses weapon systems, networks, and the underlying IT architectures that connect sensors, systems, weapons, and personnel. It also covers within its ambit applications which facilitate all these elements to function together as an integrated whole.

AI-Enabled Military Systems/ Military AI Systems

The terms “AI-enabled military systems” and “military AI systems” have been used synonymously throughout the text.