Geneva, Brussels, DC & beyond

INHR

PROPOSED MODEL PRACTICES IN TEST, EVALUATION, VALIDATION & VERIFICATION (TEVV) OF ARTIFICIALLY INTELLIGENCE-ENABLED MILITARY SYSTEMS*

August 2023

ADVOCACY DEVELOPMENT TECHNOLOGY

phone +41 79 763 73 43
email contact@INHR.org
website www.INHR.org

CONSULTATION DRAFT – NOT FINAL

This publication presents the results of the workshop held in Copenhagen, August 21–23, 2023 and sponsored by CNAS, INHR and the Royal Danish Defense College.

Table of Contents

Military AI TEVV: Need for a Global Dialogue

Integration of artificial intelligence (AI) technologies into military systems is gaining momentum worldwide and is likely to accelerate globally. Drawing from experiences in the private sector, we anticipate that all nations' governments and military forces will encounter numerous challenges when incorporating AI, including nations who are designing and deploying AI-enabled military systems, those modifying existing systems to incorporate components involving AI and those considering purchasing or otherwise interacting with such systems. For all such nations, developing a responsible, cost-benefit approach to AI regulation that considers issues arising across the life-cycle of AI-enabled military systems is important.

To promote responsible AI in the military, the international community should discuss and learn more about principles and model practices of AI Assurance – which encompasses AI TEVV and other essential components related to the safe, lawful, and ethical employment of AI-enabled military systems. Such a dialogue can contribute to mitigation of the risk of AI-enabled military systems, make those systems safer and more secure, promote compliance with international humanitarian law, and offer positive models for international cooperation. INHR has convened such a dialogue since 2018, with participation of Track II experts from Asia, Latin America, Europe and North America, including leading nations involved in design, use and deployment of AI. [1]

[1]Edelman, R. David, Li Peng, O'Sullivan, Barry, Panwar, RS, Scharre, Paul, Shanahan, John NT, Yde, Iben, Zhu Qichao, Richardson, Eric N., "Code of Conduct for AI Enabled Military Systems." Code of Conduct on Artificial Intelligence in Military Systems, HD Centre website, 18 August 2021, https://hdcentre.org/insights/ai-code- of-conduct/.

Unique Factors About AI Military Testing, Evaluation, Validation and Verification

In light of the importance of considering better regulation of AI-enabled military systems, INHR has, with the support of the Center for A New American Security, focused its international dialogue in 2023 on the testing and evaluation of AI-enabled military systems. One expert in our dialogues has posited that approximately 70 percent of the TEVV practices that militaries have developed, tested, and deployed in the past for traditional hardware-centric weapon systems are also applicable to AI-enabled military systems. But the fact that the remaining 30 percent of practices where TEVV of AI-enabled military systems is different, unique to AI, or requires consideration of AI's unique nature requires a fresh approach.

The model practices which follow are intended to provide practical guidance to states and other stakeholders considering the responsible use of AI-enabled military systems as they design new approaches for this type of TEVV of AI-enabled military systems.

Among key recommended practices and considerations unique to AI systems are:

- Continuous testing and monitoring throughout the entire AI lifecycle, from design through sustainment. AI T&E requires a deeper level of continuous technical integration among designers, testers, and end-users. With AI-enabled systems, testing is never over.
- Emphasis on iterative and incremental software development, employing Agile development approaches and adaptive T&E principles, rather than linear and sequential software Waterfall development.
- The centrality of data, including the potential for skewed, corrupted, or incomplete datasets, as well as high-end compute.
- The necessity for continuous data-based learning capabilities.
- Challenges associated with domain adaptation for AI-enabled systems.
- Probabilistic or statistically predictable (non-deterministic) behaviour.

- Effects and risks of dedicated adversarial attacks against AI models.
- Challenges in achieving AI explainability and auditability.
- Implementation of continuous integration/continuous delivery (CI/CD) for fielded AI-enabled systems.
- Importance of adding instrumentation to fielded AI-enabled systems to monitor their performance over time, as performance evolves with continuous learning.
- Unexpected and unanticipated failure modes.

The complexity of AI TEVV is amplified by the future 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 behaviors when multiple AI-enabled systems interact across weapon systems, command and control architectures, ISR platforms and new domains, such as cyber networks.

For these reasons, it is vital for all nations to determine the necessary and sufficient level of T&E throughout the entire lifecycle of AI-enabled systems, ensuring the delivery of effective, suitable, reliable, predictable, sustainable, secure, safe, trustworthy, and resilient capabilities.

The 22 AI TEVV Model Practices appended below address the major stages of the AI lifecycle, along with considerations related to international cooperation, standards setting, investigation of accidents, and AI confidence building between states. Our objective is to foster a global dialogue on this critical topic and reach international consensus among key states, while respecting the expectation that military forces will not disclose sensitive details about their respective AI TEVV processes and procedures.

The 22 AI TEVV Model Practices

I	Design and Development	
	A	AI TEVV 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. Measures should be taken to assure custody, provenance, and quality of training, testing, and validation data used.
	B	Choices of 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 TEVV process with clear performance indicators and evaluation metrics
	C	The design and development process for AI systems should incorporate TEVV requirements from the beginning. AI TEVV must account for the differences between traditional software and AI, as well as for algorithmic and operational testing.
	D	TEVV of AI systems should be viewed as a continual process. It should occur before and after a system is deployed and until a system's retirement. While TEVV for periodic AI model updates may differ in depth and breadth from TEVV prior to initial deployment of an AI-enabled system, design plans must account for continual TEVV (as part of continuous integration/ delivery).
	E	AI TEVV should include testing under real-world conditions with due regard for the resilience and robustness of the system, rollback/failsafe modes, error identification and correction.
	F	TEVV plans should specify when simulation will be used to test AI systems and how this simulated testing will be validated. For each system, it is necessary to be clear at what level of fidelity the behaviours of such system need to be tested through simulation.
	G	TEVV plans should include how testing results and system performance will be communicated to all relevant stakeholders. TEVV processes should support informed oversight by appropriate military and civilian leadership.
	H	Human-system integration and/or human-machine teaming should be considered as an integral component of TEVV design. A key aspect of TEVV is determining the ability of humans to supervise AI systems under operational conditions.
	I	Given the development of generative AI, special attention should be paid in TEVV of factors arising from foundation models and any significant updates, including the process of communication of updates and preservation of data integrity in the update process.
	J	TEVV requirements should be designed to ensure that it is possible to evaluate system compliance with relevant legal requirements, including IHL.

The 22 AI TEVV Model Practices

II	Deployment	
	K	TEVV should assess not only the performance of components and subsystems of an AI-enabled military 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.
	L	TEVV plans and systems documentation should identify a process by which, prior to deployment of an AI system to a new operational context, hazards are identified, analysed, and remediated.
	M	TEVV plans should identify high risk and catastrophic errors that could occur during operations and how these may be detected and remediated. It should also identify how unanticipated errors will be handled and 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.
	N	TEVV plans should establish how to evaluate if deployed AI systems continue to meet their performance goals; particular consideration should be given to systems that continue to learn while deployed. Appropriate corrective actions should be taken if systems do not meet these goals.
	O	Collection, management, assessment, and use of data throughout a system's lifecycle, including during 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. TEVV 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 AI TEVV Model Practices

III	Standards, Incidents and Confidence-Building	
	P	As governments work to develop and strengthen their AI TEVV 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, 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 AI on technical and/or governance issues that impact TEVV.
	R	Governments addressing TEVV of AI-enabled military systems should engage in dialogue to learn from each other, including about TEVV standards, TEVV'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 TEVV in the AI system lifecycle and to include technical, military, political and legal experts in those training programs.
	S	As part of continuous TEVV over a system's lifecycle, governments and international agencies should consider establishing standards for investigation and remediation of "incidents" that occur from the use of military AI. 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 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.
	T	To promote transparency, mutual understanding, and consistent best practice, states should publicly release aspects of their processes and approaches to TEVV 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 doctrine 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; and processes to integrate AI risk into overall consideration of system and system-of-systems risks.
	U	As governments adopt TEVV 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 to ensure sufficient TEVV before AI technology is fielded.

*As defined in the related Code of Conduct, we consider AI-enabled military systems to focus on weapons systems; intelligence, surveillance and reconnaissance; targeting; decision-support; and command and control systems. Logistical and other administrative support systems for functions ranging from personnel to supply and military medicine are important to military AI applications but are not the focus of this Guide.

CONSULTATION DRAFT – NOT FINAL

phone +41 79 763 73 43
email contact@INHR.org
website www.INHR.org