AI MILITARY MODEL PRACTICES FOR TEVV

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. 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 TEVV 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 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. As operational environments are more complex, uncertain, and less widely understood, militaries should ensure TEVV plans account for the entire AI lifecycle, to include sustainment. AI TEVV requires continual integration among developers, testers, and military practitioners to achieve predictable and reliable test outcomes.
	D	TEVV 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 TEVV for periodic AI model updates and updates necessitated on operational re-deployment or changing operational conditions 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/continuous deployment or delivery).
	E	AI TEVV 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.
	F	TEVV 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	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 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 TEVV design. A key aspect of TEVV 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 complete 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 TEVV 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 control systems are confined within negligible limits, which must be clearly specified and assessed during testing.
	J	TEVV 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.
II	**Deployment**	
	K	TEVV should assess not only the performance of components and subsystems of an AIenabled 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 network centric environment.
	L	TEVV 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	TEVV 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. TEVV 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	TEVV 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. TEVV plans must ensure that special care is taken to minimise the potential negative consequences of online learning, especially in highrisk 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. 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.
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/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 TEVV.
	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 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 military 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 "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.
	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; 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 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: the idea that introducing a new product or process whose ultimate effects are disputed or unknown should be approached using caution, pause, and review.