MAPS DIALOGUES

READOUT — SECOND 2026 MAPS DIALOGUES

# Before and After Use: A *Lifecycle* View of Military AI

*From Pre-design to Decommissioning*


DATE
28 May 2026

FORMAT
Webinar



AUTHOR
Ulysse Richard

CONTRIBUTORS
Ariel Conn · Vincent Boulanin · Heather Frase · Seb Matthews · Jonathan Kwik



United Nations Office for Disarmament Affairs | Ministry of Foreign Affairs

## CONTEXT

This is a summary of the second 2026 Military AI, Peace & Security (MAPS) Dialogue, a webinar series from the UN Office for Disarmament Affairs, supported by the Republic of Korea. The series aims to foster inclusive multilateral dialogue on military AI applications. More on the event is available here.

This second webinar of the 2026 MAPS Dialogues, *Before and After Use: A Lifecycle View of Military AI from Pre-design to Decommissioning*, examined the eleven lifecycle stages set out in General Assembly resolution 80/58 (pre-design, design, development, evaluation, testing, deployment, use, sale, procurement, operations, and decommissioning) and the analytical utility of the lifecycle framework in preparation for the lead-up to the UN informal exchanges on AI in the military domain (Geneva, 15–17 June 2026). The discussion was organized into three parts: a working map of the lifecycle as a whole, the questions that arise left of use (pre-design through evaluation and testing), and the questions that arise right of use (monitoring, repurposing, and decommissioning). The event was open to representatives of States, international and regional organizations, civil society, the scientific community, and industry.

The panel featured Ariel Conn (Consultant & Instructor, University of Colorado Boulder), Vincent Boulanin (Senior Researcher and Director of the Governance of AI Programme, Stockholm International Peace Research Institute), Heather Frase (Senior Advisor for Testing and Evaluation of AI, Virginia Tech National Security Institute), Seb Matthews, and Jonathan Kwik (Researcher, T.M.C. Asser Institute), and was moderated by Ulysse Richard (UN Office for Disarmament Affairs).

## PANEL



**Ariel Conn**
University of Colorado Boulder



**Vincent Boulanin**
SIPRI



**Heather Frase**
Virginia Tech National Security Institute



**Seb Matthews**
Industry leader



**Jonathan Kwik**
T.M.C. Asser Institute

Moderated by **Ulysse Richard**, UN Office for Disarmament Affairs

# 01 The lifecycle framework provides a map of where decisions about lawful and safe use accumulate over time

Panelists framed the eleven stages of UNGA resolution 80/58 less as a sequence to be followed than as an analytical framework for navigating the complexity of military AI. Several argued that working backward from a concrete use case makes the questions that prove intractable at the moment of use (whether a system can be used lawfully, how responsibility for an outcome can be located) tractable upstream. In the discussion, explicit comparison was drawn with how the lifecycle is already used in civil aviation and conventional weapons, where qualification and regulation permeate every stage, and where post-incident investigation reconstructs the whole lifecycle rather than only the moment of use. Panelists argued that the same logic applies to military AI, with the additional difficulty that AI systems continue to change after fielding.

For diplomats, the panel argued, the value of the framework lies less in its description of stages and more in mapping where intervention is possible: where decisions are made, documented, reviewed, and, if necessary, halted. Panelists noted that this leverage map operates at the national level, which limits the effectiveness of overly prescriptive international guidance.

Panelists also distinguished the discrete lifecycle stages from the set of activities that recur across them. These include repeated legal, ethical, and policy evaluations to continually ensure compliance and responsibility; ongoing transfers of responsibility, accountability, and knowledge as systems pass between stakeholders; continuous attention to human factors, such as recurrent training, operator education, and interface adjustments to maintain proper system use; regular technical assessments and monitoring to identify issues as systems evolve; and ongoing risk assessments to adapt to emerging challenges. Panelists argued that the international debate often focuses on individual stages and neglects recurring activities, where much of the sustained work of responsible use takes place.

# 02 By the moment of use, consequential decisions about a military AI system have already been made

The panel described lawful and safe use of military AI as a collective endeavor, with responsibility for it accumulating from the point of design.

The example developed throughout the discussion was a commander engaging a target with the help of an AI decision-support system. By the time the system reaches the commander's hands, panelists argued, developers have set the architecture and selected the data; data labelers have influenced what the system perceives; testers and legal advisors have defined the bounds of safe and lawful use; and trainers have prepared the operator to read the system's outputs critically and recognize its limits. A lack of care at any of those upstream points exposes the commander to the risk of unwittingly conducting an unlawful or otherwise undesirable attack. One panelist framed the implication as a difference between the courtroom and the design room: by the time a system is in operational hands, the most consequential decisions have already been made, and they are not visible or easily exposed at the moment of use.

Panelists drew several practical consequences from this reframing. Operators can be brought into the design phase, not just the deployment phase, so that interface and explainability decisions are informed by the people who will rely on them under pressure. Procurement officers and legal advisors can pre-publish the requirements an AI system must meet to pass a later legal review, closing a long-standing disconnect between industry and legal reviewers. The ability for humans to say no, to pause development, procurement, or use when concerns arise, was framed as a feature of human agency that should be designed in early stages. The panel returned to this point more than once: AI is not always the right solution to a given problem, and the option to decline a particular application should remain available.

## 03 The existing assurance toolkit is valuable, but its one-off logic must be revisited

The instruments that operate across the lifecycle (testing and evaluation, legal review under Article 36, procurement) were described as adapted from decades of work on complex systems, including weapons, aviation, and pharmaceuticals. Panelists emphasized that these instruments should be retained and adapted rather than discarded. What AI strains, they argued, is the set of assumptions on which the instruments rest.

Three such assumptions were introduced:

> The first is **currency of evidence**: traditional testing assumes the system under evaluation is materially the same as the system fielded later. AI systems date quickly and can degrade invisibly through online learning, memory retrieval, or model updates. The legal analog is equally serious. Article 36 of Additional Protocol I was framed for a discrete artifact, whereas AI systems can shift after review in two distinct ways: factually (through modifications, drift, or changes in operating environment or mission) and epistemically (the system itself unchanged, but understanding of it shifts as unknown unknowns emerge, since testing can never be exhaustive).
>
> The second is **diagnosability**: AI failures, particularly in generative systems, may not reproduce with the same inputs, complicating the forensic record on which accountability traditionally relies.
>
> The third is **enumerability of the action space**: complex systems with effectively unbounded action spaces have been testable in the past, but the bounding knowledge that made them testable accumulated over decades alongside incremental changes. AI advances faster than the knowledge base for testing and evaluation accumulates, and certain architectural choices (e.g., reinforcement learning, agentic systems) further expand the action space.

Panelists differed on how comprehensive testing can be in anticipating operational scenarios and risks. One panelist argued that, even with creative agents and large action spaces, specific unlawful behaviors can be anticipated and screened for during review by examining the signals a system can transmit, the training paradigm it uses, whether it is incentivized toward particular tactics, and whether red teaming surfaces them. They cited recent work on whether reinforcement learning systems could independently discover perfidious tactics (for example, masking themselves as protected medical convoys to avoid engagement) as an example of a case in which structured testing can credibly intercept a class of unlawful behavior before fielding.

On Article 36 specifically, a panelist argued for a relatively maximalist interpretation of what should trigger a fresh review, on the grounds that even modest modifications to AI systems can produce non-intuitive shifts in behavior that warrant re-examination. The panelist acknowledged that not all states share that view and that, in practice, what counts as a triggering change remains unresolved. The same panelist also distinguished what current law can mandate as review from what should be sustained to go beyond the law as assessment: a continuous monitoring obligation, supported by reporting across the fleet and upward through the chain of command, even where the law does not formally require it.

## 04 Governance is thinnest at the handoffs: between actors, between stages, and across borders

If the lifecycle toolkit's assumption strains are the conceptual problem, the operational form of the same problem is the handoff. Panelists identified four handoffs where governance is currently weakest.

**The process from pre-procurement to procurement lacks supervision.** Suppliers shape design, data, and scope choices before requirements are formally written and before any contract exists. The most rigid governance instrument during that period is often the non-disclosure agreement, and non-trivial decisions are made by people not bound to a structured process. This is also the period when legal review requirements could most usefully be published, so that industry and the review office work to a shared specification.

**Supplier visibility decreases from development to deployment.** Once a system moves to the operator, the supplier often loses access to behavior and telemetry that inform improvement and characterize failure modes. What returns through the contractual channel tends to be filtered through administrative staff rather than technical actors and to sit at the interface level (usability, feature requests, edge cases) rather than at the level of operational outcomes. A panelist noted that the information that would most improve an AI system over time, namely, performance in conditions outside its test distribution, is precisely the information that is hardest to recover because it is classified or operational. Forward-deployed engineering arrangements can close part of the gap, but only part.

**The current lifecycle stages do not adequately capture the handoff between system versions.** AI systems receive updates on time scales that bear no relation to the decade-long horizons of major military procurement, and whether a model update constitutes a new capability that needs to re-enter the framework at zero remains unresolved in most national governance frameworks.

**Cross-border sales are not adequately addressed and come with few precedents to draw on.** With traditional defense systems, hardware transfers are governed by long-standing export rules. With AI, what transfers is almost impossible to bound, from model weights and training data provenance to inference infrastructure (which globally depends on a small number of hyperscale providers) and the supplier's ongoing ability to push updates.

# 05 After deployment, the limits of the existing toolkit and the missing collective infrastructure

Focusing on the "right of use" stages related to post-operation and decommissioning, panelists highlighted that the trigger for re-evaluation remains unresolved. While software updates are a familiar trigger, distribution drift, changes in operator use, novel failure modes, and unchanged systems whose behavior is understood differently after operational exposure can all justify reassessment. One panelist proposed inverting the usual framing: rather than ask when an update warrants a fresh review, ask under what conditions positive evidence supports the conclusion that no further review is needed. The legal framework, panelists argued, rests on the obligation of constant care under international humanitarian law: not a single review, but a continuous duty of vigilance, even where the law cannot reasonably mandate a formal re-review for every weekly update.

Several panelists identified this infrastructure as largely absent. Most militaries lack a shared institutional understanding of how military AI systems fail, cross-organizational catalogs of failure modes, and standardized incident reporting that would allow a state or an alliance to recognize patterns across deployments. Panelists argued that building that infrastructure would feed back into pre-deployment evaluation by tightening the bounds on the test space. A panelist also noted a recursion problem in using AI tools to evaluate AI systems: automating evaluation pushes toward AI evaluators that themselves require evaluation, which must ultimately terminate in human judgment or a deterministic validation method.

Panelists offered different framings on the question of whether civilian and military testing and evaluation differ. One argued that the two are conceptually the same but serve different missions, with different users and varying risk-acceptance levels (with civilian applications often accepting more risk than military ones), and that the information flow between the two test communities is currently too limited. Another argued that the more pressing question is the difference between the military use of these technologies and their use in law enforcement, border control, and civil unrest response, where the operational and legal contexts shift substantially. Both views point toward testing and evaluation as an under-attended risk surface that sits across the civilian-military boundary, and that current protective frameworks were not designed for.

Panelists described the end of a system's life as an often-overlooked part of the lifecycle. They argued that decommissioning is fundamentally a planning and documentation problem: deciding which artifacts (model weights, training data, test data, architecture, persistent memory) need to be controlled, by whom, and for how long, and that those decisions have to be made before the system is built, not when it leaves service. Decommissioning also has system-of-systems consequences: a retired system may leave a hole that other operationally networked systems will need to be patched.

# 06 Revisit existing instruments in continuous and contractual form

Panelists converged on the proposition that the work ahead does not principally require new AI-specific measures. The instruments States already have (legal review, testing and evaluation, procurement, training, doctrine, incident reporting) can be preserved, adapted, and reinforced for systems that change continuously, fail invisibly, and depend on actors well outside the chain of command. The lifecycle framework, panelists argued, is the means by which those adaptations can be located.

Panelists identified several practical levers for delegations preparing for the informal exchanges on AI in the military domain:

→ **Move decisions currently made at the moment of use earlier in the lifecycle:** pre-publishing the requirements an AI system must meet to pass a legal review, involving operators in design, and involving operational testers in pre-deployment evaluation with the data hooks and traces they will need after fielding.

→ **Convert one-shot processes into continuous ones:** pairing Article 36 review with a constant-care assessment obligation and shifting procurement from a linear gate to an iterative model in which each decision point carries fresh review obligations.

→ **Make informal arrangements contractual:** structured feedback from operational use back to the supplier should become a contractual requirement rather than an aspirational ask, and post-deployment changes should be governed by in-process review rather than left to ad hoc determination.

→ **Preserve the institutional ability to say "no":** Organizations should build "stop mechanisms" that allow humans to halt development, procurement, or use early in a project's lifecycle, rather than constructing them after a project has acquired momentum.

WHAT'S NEXT · WEBINAR III

## Working the Disagreements: From Tensions to *Traction* in Military AI

10 June 2026 · 09:00–10:30 New York / 15:00–16:30 Geneva (90 min) · Virtual panel

The third and final 2026 webinar stress-tests where progress is possible across three areas of tension (substance, process, and access), feeding directly into the UN informal exchanges on AI in the military domain (15–17 June).

REGISTER

BIT.LY/MAPSDIALOGUES3