MAPS DIALOGUES

READOUT — FIRST 2026 MAPS DIALOGUES

A Shared *Vocabulary* for Military AI

Bridging technical understanding and policy needs

DATE
7 May 2026

FORMAT
Webinar

AUTHOR
Ulysse Richard

United Nations Office for Disarmament Affairs | Ministry of Foreign Affairs Republic of Korea

CONTEXT

The following is a non-exhaustive summary of key elements emerging from the discussion held during the first session of the 2026 edition of the Military AI, Peace & Security (MAPS) Dialogues, a webinar series convened by the UN Office for Disarmament Affairs, with the support of the Republic of Korea, to foster inclusive multilateral dialogue on military applications of AI.

This first webinar of the 2026 MAPS Dialogues, *A Shared Vocabulary for Military AI: Bridging Technical Understanding and Policy Needs,* focused on the misalignment between technical and policy communities in the use of key terms, the meaning in practice of contested concepts such as trust, accuracy, predictability, transparency, explainability, and human control, as well as the pathways available at the national and multilateral levels to support shared understanding ahead of the informal exchanges on AI in the military domain (Geneva, 15–17 June 2026). The event was open to representatives of States, international and regional organizations, civil society, the scientific community, and industry.

The session was opened by H.E. Ambassador Jongin Bae, Deputy Permanent Representative of the Republic of Korea to the United Nations, and convened over 160 participants from across the diplomatic, technical, academic, and civil society communities. The panel featured Zena Assaad (Australian National University), Heidy Khlaaf (AI Now Institute), and Paul Lushenko (Cornell University), and was moderated by Ulysse Richard (UN Office for Disarmament Affairs).

PANEL

Zena Assaad

Australian National University

Heidy Khlaaf

AI Now Institute

Paul Lushenko

Cornell University

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

01 DIAGNOSING

Where vocabulary diverges between technical and policy communities

Discussions identified recurring patterns of misalignment between technical and policy communities working on AI in the military domain, as well as several problems, such as: communities using the same words to mean different things, different words to refer to the same underlying concept or valuing trade-offs differently.

Panelists highlighted several sources of divergence. Some argued that technical disciplines define terms without subjectivity, deliberately narrowing each definition to a single interpretation so that the term can be measured or quantified, while other domains, including law, doctrine, ethics, and diplomacy, may retain subjective interpretation by design. Even within the technical AI community alone, multiple definitions coexist: some panelists observed that traditional safety and security engineering terms, including red teaming and safety itself, have been repurposed inside parts of the AI community to mean something other than their original definitions, compounding the confusion that reaches policy discussions.

The discussions surfaced a range of specific terms were flagged as carrying different meanings across communities, including *trust, reliability, accuracy, predictability, autonomy, AI-enabled warfare, human-machine teaming, meaningful human control,* and *appropriate human judgment.* Panelists discussed how each of these is anchored differently depending on the speaker's frame of reference. As one example, a panelist noted that the United States Department of Defense Directive 3000.09 is based on appropriate human judgment, while the United States' allies, partners, and multilateral institutions may not adopt the same definition.

Some expressed concern about the downstream consequences of these mismatches. Panelists converged on the perception of accuracy built on a broader rather than a narrower definition can lead to over-trust in AI-enabled systems and, in turn, to compressed timeframes for human decision-making in operational settings — including in targeting. It was also noted that under-trust is a real and parallel concern, particularly in hierarchical military environments where junior leaders may rely on observation of peers and superiors rather than on independent assessment of a system's capabilities, producing what one panelist described as "trust by proxy."

On the question of civilian harm, while some panelists pointed to civilian casualty rates in current operations as a manifestation of these breakdowns, another panelist argued that the pattern is more accurately read as the outsourcing of human control rather than a property of the systems themselves, citing what they described as a historically unprecedented reduction in civilian casualties achieved by certain militaries using large, armed and networked drones over the past two decades. At the policy level, it was noted that multilateral discussions can produce text in which there is broad agreement on the words but persistent disagreement on what those words require in practice.

The discussion also touched on an accountability gap between commercial AI vendors and States. A panelist argued that vendors caveat their evaluations by placing responsibility for use on the deploying military, while policymakers may proceed under the assumption that legal and ethical considerations have been engineered into the system. Experts also noted that vendor metrics are not always scientifically confirmed or independently verified, and that vendors are under no obligation to consider international humanitarian law (IHL) in their evaluation regimes.

02 UNPACKING
How key concepts work in practice

Trust and reliability were discussed in detail. Panelists distinguished between technical reliability — typically defined as a measurable property of a system under specified conditions — and operational trust, characterized as contextual, multi-stakeholder, and dependent on perceived effectiveness, civilian-harm risk mitigation, and oversight mechanisms, both domestically and internationally. Two panelists offered their working definition of trust: *"confidence in the reliability of a system when it is used in its intended operation of use,"* and *"perceptions that a system will reliably perform as expected."* Some panelists emphasized that trust and reliability should not be conflated, given that technical evaluations alone do not determine whether a system is used responsibly.

The relationship between accuracy and predictability received specific attention. One panelist noted that accuracy, as used in AI literature, refers to the percentage of correct predictions made by a model relative to the test set on which it was evaluated, a metric whose usefulness is inherently restricted to the scenarios reflected in the testing and training data. By contrast, predictability, central to ongoing discussions on lethal autonomous weapon systems, was described as a broader operational property requiring the ability to anticipate how a weapon will function across the full range of expected circumstances of use. A panelist expressed concern that treating accuracy and predictability as interchangeable risks anchoring policy frames around metrics that cannot be substantiated under operational conditions, particularly given that neural networks do not generalize beyond their training data.

The discussion unpacked the terms transparency, explainability, and traceability as serving distinct functions for distinct stakeholders. Panelists discussed transparency as a property of the system as a whole — concerning what is in it, how it was built, and how it behaves — and operating differently in military contexts than in civilian ones, given national security constraints. A speaker noted that military transparency should focus on enabling specific institutions to hold developers and users accountable, rather than mirroring civilian-style public disclosure. Explainability, defined as the ability to understand why a system produced a specific output, was characterized as an unsolved open problem in AI: outside classic symbolic systems such as decision trees, panelists noted that complex models, including neural networks do not provide explanations of their behavior. Traceability, defined as the capacity to follow the audit trail back through data, model architecture, and training, was identified as a precondition for evaluating failure modes and the modes of operation for which a system is suitable.

Panelists also discussed the trade-off between system performance and explainability. One speaker noted that increasing the capability of an AI model tends to reduce its explainability without necessarily improving its accuracy or precision, referencing research suggesting that smaller, purpose-built models can outperform larger general-purpose models on tasks requiring domain expertise, including military tasks. Panelists also emphasized that the appropriate trade-off depends on the specific application and context; surveillance and targeting were cited as cases where the consequences of system fallibility differ substantially.

Several panelists raised what was described as a prior question, whether AI is the right tool for a given task in the first place. The view was expressed that most discussions often start from the premise that AI is more capable than it is, and treat its integration into military applications as a given inevitability, leaving little space for the conclusion that a particular AI approach may simply be unsuitable for a particular function. Multiple panelists suggested that fitness for purpose should be assessed before evaluation criteria are debated, and that the policy community should preserve the option not to employ specific applications in the military domain.

03 ACTING
Pathways at the national, interface, and multilateral levels

Panelists addressed pathways for action across three interrelated layers — national, interface between communities, and multilateral — and emphasized that coordinated work is needed across them.

A National approaches

Panelists addressed what States can require today through national procurement, testing, and evaluation processes. Some called for a deliberate slowing of the procurement cycle, arguing that what was characterized as a *"fabricated urgency"* associated with the current AI arms race is producing the shortcutting of established testing, evaluation, verification, and validation (TEVV) practices. The view was expressed that systems deployed without these practices lack technical integrity in what are essentially safety-critical contexts, and that requiring more time for TEVV is a national-level lever rather than a multilateral one.

It was also noted that existing military safety and security standards already provide thresholds that AI systems should be required to meet, and that, even where methodologies need updating to accommodate AI, practices such as field testing, modeling and simulation, war games, and independent technical assessment should be preserved. It was suggested that, where AI systems cannot meet traditional thresholds, this should be taken as an indication that the systems are not suitable for that application, rather than a reason to lower the threshold or substitute alternative testing frameworks such as general benchmarking. Some panelists indicated the need to discuss acceptable risk thresholds for AI-enabled systems, drawing on the systems-engineering principle of as low as reasonably practicable (ALARP) and noting that an articulated answer for risk thresholds in the military domain is not yet available.

B Technical-policy interface

A recurring theme was the importance of the interface between technical and policy communities. Panelists argued that policymakers cannot ask the right technical questions in isolation, and that technical experts need to be present in policy discussions rather than only consulted reactively. Several panelists emphasized the importance of independence in technical assessment, noting that verification and validation conversations are increasingly driven by the very companies developing and selling the systems being evaluated, and that domain expertise, in safety engineering, security engineering, and the relevant operational fields, should not be treated as subordinate to AI expertise.

Examples of institutional mechanisms for operationalizing this interface were discussed, including advisory bodies that bring together technical and operational expertise across domains and convene specialists for time-bound consultation on specific questions. It was also noted that the asymmetry between expectations placed on policymakers, who are expected to upskill on technical questions, and on technical communities, who are not always asked to engage with policy processes, should be addressed in the design of these mechanisms.

C Multilateral approaches

At the multilateral level, panelists discussed the need for a technically informed and endorsed taxonomy of the key concepts in current discussions, including human oversight, AI-enabled warfare, and human-machine teaming. Panelists suggested that each of these concepts should be problematized by function, domain, and scale, given that the operational implications vary significantly across these axes. Some panelists pointed to existing variation in oversight configurations as illustrative of the level of specificity that policy frameworks may need to engage with to support both measurability and accountability.

Several panelists called for a shared policy framework that States could operationalize, noting that consensus on broad principles often coexists with significant fragmentation during their implication, including across regional groupings, and that this fragmentation has consequences for interoperability and for civilian-harm mitigation in multinational operations. One panelist also suggested that use cases organized by function, domain, and scale would be of particular value to States that do not currently have the technological or financial capacity to acquire AI-enabled systems but expect to in the future.

Finally, one panelist expressed that multilateral processes should aim for a more homogeneous approach across forums, connecting existing conversations rather than starting parallel initiatives that re-explain the same terminology in different contexts. Panelists noted that a more coordinated approach would reduce duplication and help focus attention on the specific policy questions where shared understanding is most needed in the lead-up to the UN informal exchanges in Geneva.

WHAT'S NEXT

The next MAPS Dialogues will focus on the life cycle of AI systems.

PRE-REGISTER →