AISE MARKERS SERIES

Benchmark II: Technology

Insights from the Global Conference on AI, Security and Ethics 2025

YASMIN AFINA · JAN HENDRIK MANNSPERGER

1. Introduction

1.1. Context

On 27–28 March 2025, UNIDIR organized its inaugural Global Conference on Artificial Intelligence, Security and Ethics (AISE25), hosted in the Palais des Nations, Geneva. Led by UNIDIR's Security and Technology Programme, the conference provided an agile response to rapid advances in artificial intelligence (AI), which have put this technology at the forefront of today's global policy discussions. AISE25 was held as policymakers and regulators worldwide increasingly recognized the urgency of developing shared understandings, norms and regulations that can transcend national borders and individual interests, including in the context of peace and security.

The conference sought to provide a unique forum for engagement between the multilateral ecosystem and the wider multi-stakeholder community, including academic experts, civil society organizations, industry representatives and research laboratories interested in the governance of AI in peace and security. By jointly analysing and addressing the complex implications of AI for national, regional and global security and resilience, AISE25 enriched dialogue between participants and exposed them to the latest research in the field. This opportunity to exchange and consolidate views on AI in the military domain was timely, with the United Nations General Assembly having requested the views of Member States and other stakeholders as a means of informing discussions during its Eightieth Session, in September 2025 – the deadline for which came just weeks after the conference.[1]

The conference programme was designed to build on the work undertaken as part of UNIDIR's Roundtable for AI, Security and Ethics (RAISE), a multi-year project on multi-stakeholder engagement in this space launched with the support of Microsoft.[2] Specifically, the conference's agenda was primarily organized around the six priority themes identified at the inaugural edition of RAISE, which took place in Bellagio, Italy, in March 2024:[3]

1. Knowledge and capacity-building
2. Trust-building
3. The human element
4. Data practices
5. Life cycle management
6. Addressing destabilization

Combining a series of panel discussions, presentations in the form of thematic deep-dives and lightning talks, as well as a poster exhibition, AISE25 provided a timely platform, open to all, to jointly consider and elaborate on each of these six priority themes while promoting meaningful dialogue and cooperation.

Ahead of the second edition of the AISE, in June 2026, a series of three reports – the first of the AISE Markers series – takes stock of the key takeaways from AISE25 to provide an initial basis and scaffolding for AISE26. By acting as a bridge between editions of AISE, the AISE Markers series will ensure that each conference is built on solid ground and constitutes a natural evolution from the discussions held in the previous conference.

This second report gives a structured account of where the technology stood as AISE25 took place. It provides an overview of shared understandings – particularly in the areas of dual-use transformation, life cycle management, data and assurances – and identifies areas that may subsequently serve as a baseline for AISE26. Accompanying reports will cover the state of governance and use cases.

1.2. The technical landscape in March 2025

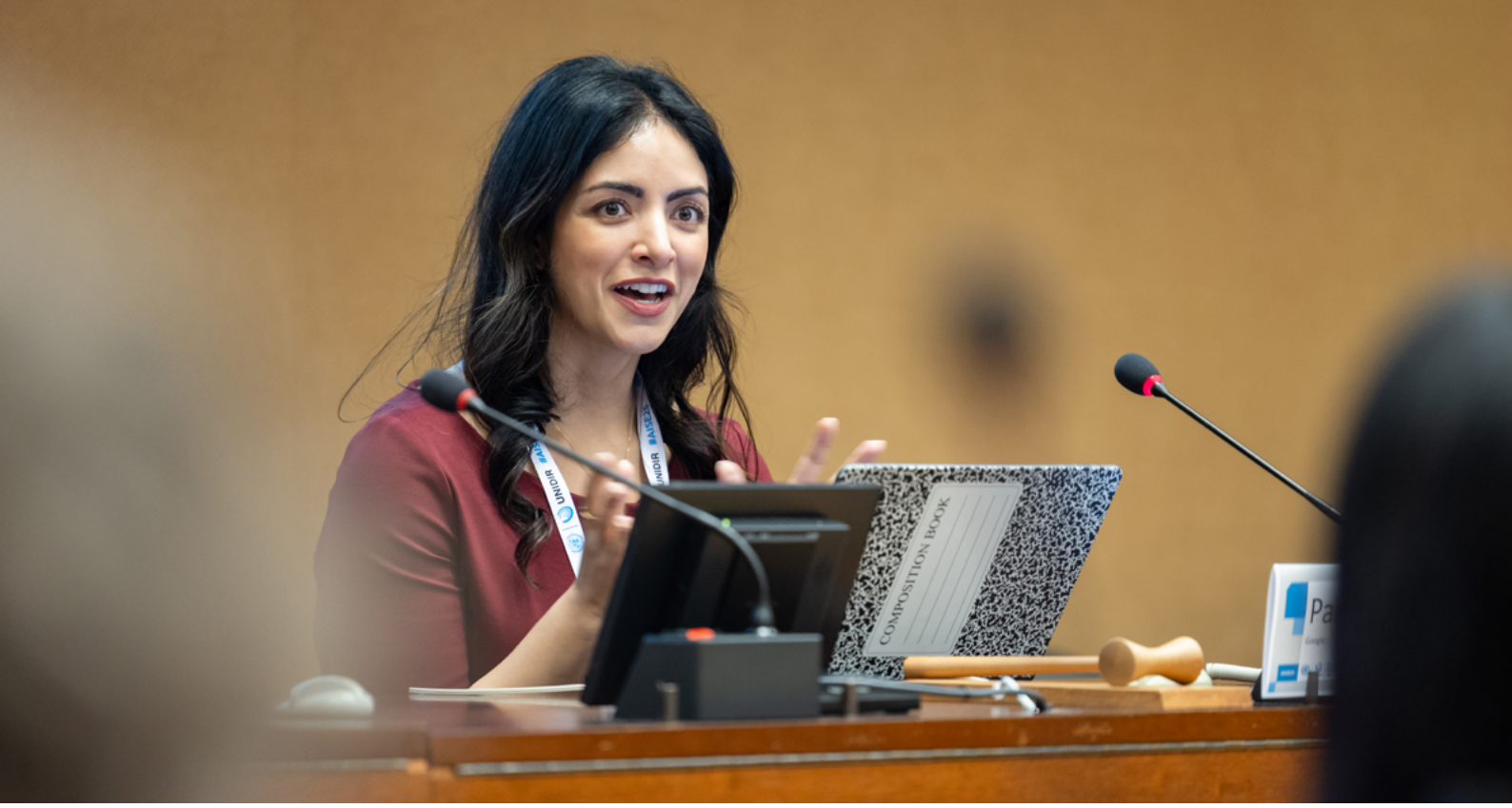
Against the backdrop of active multilateral governance debates and a rapidly evolving institutional landscape, AISE25 sought to provide grounding in what military AI technology actually is and what it currently does. At the time of the conference, AI was generally recognized as an operational reality across a range of security and defence contexts:

---

1 General Assembly, resolution 79/239, "Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security", 24 December 2024, https://docs.un.org/A/RES/79/239.

2 UNIDIR, "RAISE: The Roundtable for AI, Security and Ethics", https://unidir.org/raise/.

3 Y. Afina and G. Persi Paoli, *Governance of Artificial Intelligence in the Military Domain: A Multi-Stakeholder Perspective on Priority Areas* (Geneva: UNIDIR, 2024), https://unidir.org/publication/governance-of-artificial-intelligence-in-the-military-domain-a-multi-stakeholder-perspective-on-priority-areas/.



the question, as framed in the conference opening, was no longer whether AI would be used in military and security contexts, but how. Understanding the technical dimensions of that question – that is, what AI systems actually are, how they fail, what kinds of deployment have already occurred and what governance implications emerge – was perceived as a prerequisite for ensuring that governance efforts and subsequent frameworks could be calibrated to the actual technological landscape, rather than to broad, over-generalized, anticipated or even idealized versions of it.

Four thematic areas structure this report's account of the technological landscape: the dual-use transformation that characterizes the relationship between civilian and military AI development; data as a foundational element; the governance dimensions that emerge from the management of AI across its full life cycle; and the technical foundations of trustworthiness and accountability that any responsible deployment requires. Out of each of these areas, specific issues and challenges emerged that are not merely technical in character and nature, but also have direct implications for the governance and institutional frameworks examined in the first paper of this series, on governance.

One preliminary observation is necessary before engaging with these themes: AI, as a category, encompasses a range of technologies, techniques and applications whose governance implications differ substantially from one another. The characterization of AI as a single, monolithic object of governance increases, rather than diminishes, uncertainty and complexity. Distinct governance questions are raised by data science techniques that do not involve machine learning, optimization algorithms that have operated reliably for decades, and large language models (LLMs) whose outputs are probabilistic rather than deterministic. Governance frameworks that treat them as equivalent would, arguably, be as poorly calibrated as those that treat AI in the security domain as an entirely new field.[4]

4 As noted by Paree Zarolia (Google) in a dedicated deep-dive on day 1: "Addressing Safety and Risk in the AI Development Lifecycle".

2. The dual-use transformation

As AI capabilities grow more multi-purpose and effective across an expanding range of applications, systems initially designed and developed for civilian purposes are increasingly perceived as also offering opportunities for the military and security domains. The boundary between civilian and military domains is becoming blurred, with the continued exploitation of the inherently complex nature of these technologies and the subsequent spillover capacities. The extent to which international, regional and national governance regimes and frameworks are adequate to address this evolution is subject to much debate. In response, AISE25 mapped a spectrum of dual-use implications and risks across current AI models and highlighted the possible threats posed by rapid and unchecked transfer from civilian to military applications.

2.1. The militarization of civilian AI

The division between civilian and military AI is less a clear boundary and more a continuum, and one that is perceived as disappearing faster than governance frameworks have been able to track. A defining feature of contemporary AI developments is that the same systems that power commercial applications are available, and in many cases are actively being repurposed, for military use. The general-purpose design and nature that make AI commercially successful – that is, maximum adaptability, large training data sets and a broad spectrum of applicability – are precisely what makes AI attractive for military applications.[5]

The clearest illustration of this dynamic is not speculative but historical: the United States Department of Defense's Project Maven, launched in 2017, reportedly began by adapting a commercial image-recognition algorithm, readily available online, to the task of gathering intelligence for missions that involve uncrewed vehicles.[6] No new weapon system was involved and no fundamental redesign was required: only a redirection of purpose. Within a month of the adoption of a civilian, commercial tool, the platform had become one for warfighting. The pace of this transition establishes the magnitude of the governance challenge: regulatory and normative frameworks designed around the assumption of distinct civilian and military AI ecosystems are perceived as inadequate in an environment in which that distinction may become rapidly – and increasingly routinely – obsolete.

The mechanisms through which civilian AI becomes militarized are themselves diverse. Technologies originally developed with privacy preservation in mind (e.g., federated learning, which enables models to be trained without centralizing sensitive data) are now being used to enable frontline forces to train AI locally without exposing location data. In this way, a design feature developed to serve civilian data-protection interests becomes a capability that enhances operational security. The result is that every commercial breakthrough becomes, in principle, a military capability.

5 As noted by Yuyin Liu (Asia-Pacific Artificial Intelligence Association) in a dedicated deep-dive on day 1: "The Unseen Militarization of Civilian AI and the Rise of Algorithmic Warfare".

6 Liu.

This dynamic has direct implications for governance design. A framework oriented towards regulating AI based on what a technology is, rather than what it does and in whose hands it operates, could be systematically outpaced. It also raises questions about the private sector's role and position amid this trend: where a commercial AI system is adapted for military purposes by a state, the developer may bear no formal responsibility for that adaptation, even if the original design characteristics of the system facilitated it.[7]

Consequently, awareness of this dynamic at the national level is arguably a precondition for governance. In the Netherlands, for instance, the rapid informal uptake of publicly available LLM tools by tens of thousands of military personnel, using platforms not designed for sensitive government information, prompted the development of an isolated, Internet-disconnected equivalent. The governance lesson is structural: adoption of civilian technology by military personnel has preceded policy responses, and governance frameworks that wait for policy cycles to complete will routinely find themselves operating in an environment that has already moved on.[8]

2.2. Autonomous drones and the spatial security challenge

Of the specific technological developments examined at AISE25, the proliferation of commercially produced autonomous and semi-autonomous drones was identified as presenting a unique security and governance challenge. This challenge is not exclusively about the capabilities of state military actors, but also about the implications of widespread, low-cost access to platforms that acquire a physical presence.[9]

The commercial drone market has driven a wholesale proliferation of – and low-cost access to – flight that security architectures designed for earlier eras were not necessarily designed nor built to accommodate. The ease with which commercially produced drones, sold for recreational and hobbyist use, can be operated in or near sensitive airspace illustrates the scale of the problem: these platforms are available to almost anyone with the capacity to purchase them, regardless of intent. Legal and institutional frameworks governing civil aviation were not designed for this environment; nor, in most cases, were the physical security installations that protect sensitive facilities.

The specific technical challenge is compounded by the interaction between autonomy and the primary available countermeasure: radio frequency jamming. Jamming relies on disrupting the communication between a drone and its operator. Yet, a drone that has been trained to navigate by reference to an internally stored map of a terrain would not necessarily depend on receiving directional instructions and will continue towards its programmed destination autonomously, regardless of signal disruption. The greater the degree of autonomy, the less relevant the primary detection and disruption tool becomes: a relationship that creates direct incentives for the development and deployment of increasingly autonomous platforms.

The governance dimensions of this challenge were developed at AISE25 through the concept of spatial security: the recognition that territory,

7 As noted by Arnaud Valli (Comand AI) in day 2's opening panel on tech leaders.

8 As noted by Jeroen van der Vlugt (Ministry of Defence, Netherlands) in a dedicated deep-dive on day 1: "Building a Defence AI Strategy – The Dutch Approach".

9 As noted by Troels Boe (Advisor to the Danish Technology Ambassador to Silicon Valley) in a dedicated deep-dive on day 1: "Circumventing Security – Autonomous Drones and Security in Urban Environments".

for security purposes, must be understood as a volume rather than an area, and that security installations designed around a two-dimensional perimeter model are structurally vulnerable to threats that operate in the vertical dimension. Most contemporary security installations are designed around a two-dimensional understanding of threat (e.g., through fences, checkpoints, perimeter monitoring); the adaptation of these architectures to volumetric threats involves not only technical adjustment but institutional and legal reorganization. Civil aviation, law enforcement and military defence each operate under different legal mandates and institutional cultures, but effective spatial security governance will require cross-sector cooperation that those frameworks, currently, neither provide nor facilitate.

The broader observation that emerges from this analysis is about the pace of civilian technological development relative to the adaptation of security governance frameworks: the same commercial market dynamics that have made advanced AI capabilities widely accessible have also made small, capable, autonomous aerial platforms available at a price point that places them within reach of a wide range of actors, both state and non-state. The governance implications of this proliferation for law enforcement, for military security and for civilian infrastructure protection are not speculative; they are already being navigated in active conflict environments and, increasingly, in peacetime settings as well.

2.3. AI-enabled societal destabilization

Beyond the physical dimension, AI presents a further set of challenges in the informational and cognitive domain: the same capacities that have made AI commercially valuable – the abilities to generate, analyse and personalize content at scale – are increasingly applied to the systematic manipulation of the information environments on which individual and collective decision-making depend.[10]

The mechanism involved is not primarily the creation of false information (although it constitutes one important dimension): instead, it is the construction of an information ecosystem in which the analytical frameworks through which individuals and communities make sense of their experience are gradually reshaped. Echo chambers, filter bubbles, polarization and targeted micro-campaigns represent a suite of tools that, individually and in combination, can substantially alter the ways in which a population perceives threats, opportunities and the trustworthiness of institutions. Scaled up to a collective level, the erosion of shared epistemic foundations reduces social capital and creates conditions in which governance institutions (including those responsible for managing armed conflict and negotiating international, regional and national security arrangements) may lose their legitimacy.

The significance of AI in this context is as a force – and threat – multiplier: the capacity that AI provides for targeted content generation, profile analysis and personalized delivery does not create the phenomenon of information manipulation, but it dramatically expands the scale, speed and precision of the manipulation. The governance challenge is correspondingly amplified: frameworks designed to address information manipulation in a pre-AI environment are not calibrated to the speed, scale or specificity with which AI-enabled campaigns are, or can be, deployed.

An additional dimension of concern raised at AISE25 concerned the aggregation of AI

10 As noted by Krishnakumar Gurumurthy (Volvo Cars) in a lightning talk on day 1: “Multi-Dimensional Security Risks and Social Destabilization Concerns of AI”.

systems across defence and security applications: as multiple AI systems from different developers and different countries are integrated into a single operational architecture, questions arise about whether commanders and operators retain meaningful situational awareness and, by extension, meaningful accountability over the combined output of those systems. If command and control are exercised over a system of systems whose aggregate behaviour cannot be reliably predicted from the properties of its individual components, the attribution of responsibility for outcomes becomes accordingly difficult.[11]

3. Managing AI across the life cycle

3.1. The life cycle definition problem

Any approach to the governance of AI in defence and security must begin with reflection on what the life cycle of an AI system encompasses, and it must acknowledge that no universally shared or agreed definition of that concept currently exists. Different institutional actors (i.e., corporate entities, civil society organizations, international and regional institutions) would produce as many different accounts of the relevant stages as there are actors asked to provide them. This definitional divergence is not merely academic: it has direct practical consequences for governance.[12]

The questions of who bears responsibility for an AI system's behaviour at any given point, which testing and validation standards apply, and what constitutes appropriate governance of a system that is no longer actively deployed cannot be answered without a shared understanding of the life cycle stages to which those questions apply. In the absence of that shared understanding, governance discussions that appear to address the same questions may in fact be addressing different phases of a system's existence in potentially incompatible ways.

A further dimension of this problem concerns the distribution of ownership and responsibility across the life cycle. In most cases, different actors will be involved in the development, the procurement, the testing, the deployment and the eventual decommissioning of an AI system, and they will not necessarily operate within the same legal or institutional framework. The governance of AI across the full life cycle therefore requires not only a shared understanding of its stages, but also an account of which actors bear what responsibilities at each stage – an account that does not presently exist in a systematic manner.

3.2. Technical validation: Advantage or Achilles heel?

The deployment in security and defence contexts of an AI system that does not perform as intended, or that performs as intended under normal conditions but fails under adversarial pressure, is not a hypothetical future risk. It is a present operational reality, and technical validation is the mechanism through which it is meant to be managed. Rigorous technical validation is thus arguably not a luxury but an absolute necessity for AI systems deployed in security and defence contexts: such systems must function reliably in the most unpredictable

11 Valli.

12 As noted by Alexi Drew (International Committee of the Red Cross) in the facilitation of the life cycle management session on day 1.

and hostile environments imaginable, characterized by weather variation, seasonal change, terrain transformed by conflict, adversarial adaptation and the deliberate exploitation of system vulnerabilities.[13]

The adversarial dimension of this challenge is particularly significant. Adversarial attacks (i.e., techniques designed to exploit weaknesses in AI models) are not exclusively digital: they can be physical, creative and capable of being deployed across the full range of AI applications, from computer vision to natural language processing. The strategic purpose of such an attack is not necessarily to cause the AI system to fail outright, but to degrade confidence in its outputs: to make decision makers hesitate, to introduce uncertainty into a decision cycle and to slow down processes expected to operate at machine speed. A false positive injected into an AI-powered object-detection system feeding an indicators-and-warnings matrix may produce an intelligence assessment that propagates through allied networks before the error is identified, with consequences that may be significant in time-sensitive operational contexts.

Technical validation encompasses three distinct elements that are not interchangeable:

- Reliability testing, which validates whether a system meets its claimed performance parameters
- Robustness testing, which evaluates whether the system maintains acceptable performance when presented with inputs that differ from its training distribution
- Security testing, which assesses resistance to deliberate adversarial exploitation

A validation regime that addresses one of these elements but not the others provides only partial assurance. More fundamentally, the three elements require different expertise and different methodologies, and so governance frameworks that conflate them will produce validation processes that are incomplete by design.

A structural principle that emerged from the conference's technical discussions is the importance of integrating validation throughout the development process, rather than treating it as something to be checked at the end. The analogy from engineering is direct: bridges are not constructed and then tested to determine whether they will stand; engineering principles are applied throughout their construction such that, by the time the structure is complete, its performance characteristics are already understood. An AI system that reaches the final stages of development before substantive validation is conducted leaves almost no room to address the vulnerabilities that validation reveals. In contrast, spreading validation across design, development and deployment stages means that the final assurance review confirms what is already known, rather than uncovering what has been missed.

A corollary concern, addressed in the conference sessions on life cycle management, relates to decommissioning: governance of the disposal of an AI system requires the same systematic approach as governance of its development. A system determined to be no longer fit for operational purpose must be disposed of in a manner that accounts for the data it contains, the capabilities it embodies and the records of its past decisions; and the actors who rely on that system must be clearly informed when it is no longer available.

13 As noted by Ben Fawcett (Advai) in a dedicated deep-dive on day 1: "Advantage or Achilles Heel? How Technical Validation Can Mitigate the Major Risks of Operating AI in Defence".

3.3. The Auditable, Controllable, Transparent and Secure framework

The Auditable, Controllable, Transparent and Secure (ACTS) framework was presented at AISE25 as a practitioner-developed contribution to the operationalization of norms for responsible development and deployment of AI. It is a framework that has emerged from direct experience of conflict-forecasting systems used in policy and operational contexts. The ACTS framework describes itself as a grassroots initiative at the intersection of technology and the social sciences, intended to complement rather than replace existing governance instruments.[14]

The motivation for the ACTS framework was grounded in a specific practical finding: a conflict-forecasting system that produced measurably accurate predictions nonetheless failed to generate trust among its intended users. No amount of performance metrics altered this outcome. The inference drawn from this experience was not that the users were wrong to withhold trust, but that the absence of trust was appropriate – a design signal, rather than a user failure. A system whose workings are opaque, whose failure modes are unknown and whose outputs cannot be interrogated in accessible terms does not merit unconditional trust, regardless of its technical performance. The aspiration of the ACTS framework is to describe the conditions under which trust is warranted, rather than to assert that it should be extended on the basis of claimed performance alone.

Three cases of AI system failure were presented as motivation for the ACTS framework in order to illustrate the range of contexts in which these conditions matter. The first concerns automated flight-control systems, where hidden automation and missing safeguards fatally undermined human judgment, particularly in contexts of implicit trust. In this case, the assumption that the system was functioning correctly was not tested against available evidence. The second concerns a financial trading system, where a software update was deployed on seven of eight servers, with no automated check made to ensure complete deployment before live trading commenced. Within 45 minutes of deployment, the consequence was catastrophic. The third concerns a system for prediction of public health that initially achieved high accuracy but subsequently overestimated cases by 140 per cent. This resulted in a loss of trust and its eventual abandonment, illustrating that predictive systems require continuous evaluation, testing and monitoring to remain both accurate and reliable.

Against this empirical backdrop, the four ACTS pillars – audit, control, transparency and security – can be understood as responses to specific failure modes, rather than as abstract governance principles. Auditability requires the systematic documentation of data sources, model versions and performance metrics, and the maintenance of ongoing checks. The failure to audit a model continuously after deployment is a governance gap in its own right, not an opportunity for subsequent optimization. Controllability requires that a human actor can always intervene to halt or correct a system's operation. This includes through scenario-based exercises that test the human infrastructure surrounding the computational system, not merely the system in isolation. Transparency, in the ACTS framework, does not mean technical explainability in the conventional sense; it means the generation of

14 As noted by Alexa Timlick and Simon Polichinel von der Maase (Peace Research Institute Oslo, PRIO) in a dedicated deep-dive on day 1: "Auditable, Controllable, Transparent, Secure (ACTS) Now: Why Machine Learning Operations Must Govern AI in Critical Systems and High-Stakes Domains".

knowledge from a system that can be communicated accessibly to both technical and non-technical users. This is analogous to the dashboard of a vehicle, which presents relevant information in actionable form while appropriately withholding irrelevant detail. Security, finally, means designing systems to fail loudly and promptly when anomalies occur – that is, ensuring that both the technical and human infrastructure surrounding the system are alerted when unusual conditions are detected, rather than allowing the system to continue operating under conditions that suggest malfunction.

The ACTS framework's significance at AISE25 lay in providing a practitioner-grounded account of what operationalizing responsible AI actually requires at the level of system architecture and monitoring practice; governance frameworks articulated at the level of principles rarely supply this.

3.4. Safety and risk in AI development

The approach taken by large AI developers to managing safety and risk across the development life cycle was examined at AISE25 through a specific application domain: the development of large language models and the safety practices that accompany their development and deployment. LLMs are probabilistic systems: they predict words likely to appear next in a sequence, based on patterns learned from training data. They are not information-retrieval systems and do not produce factually reliable outputs by default; they generate statistically plausible text, which may or may not be accurate.[15]

This foundational characteristic has direct implications for governance: systems whose outputs are probabilistic rather than deterministic present accountability challenges that are different in kind from those presented by deterministic software. Moreover, frameworks that treat AI outputs as straightforwardly factual or reliable will be ill-equipped to manage those challenges. The governance question is not only what an AI system produces, but the evidentiary status of what it produces. This question varies across application contexts and has particular significance in security and defence settings, where AI outputs inform decisions with significant consequences.

The safety and risk practices described at AISE25 in this context involve several distinct mechanisms. Red teaming (i.e., the systematic attempt to elicit problematic outputs from a system through adversarial inputs) is employed both by human evaluators testing the boundaries of system policies and by generative models themselves, using AI to identify novel and unexpected failure modes. The development of exemplary outputs for specific query types provides training signal for the system, but the governance dimension of this practice lies in the judgments required to determine what constitutes an appropriate response to a sensitive or contested query. These judgments necessarily involve assessments of values and potential harm, not merely technical optimization.[16]

An empirical finding of particular governance relevance concerns the relationship between content labelling and user inference. Research reported by Googlehas found that roughly one in seven users interpret unlabelled content as content that has not been modified at all. The governance implication – sometimes described as the implied truth effect – is that labelling strategies which identify some content as AI-generated create an implicit

15 Zarolia.

16 Ibid.

signal about unlabelled content. This is taken to mean that, in an adversarial information environment, partial labelling systems create exploitable vulnerabilities that labelling alone cannot address.

Technical provenance mechanisms, such as the embedding of imperceptible but machine-detectable watermarks into AI-generated images and audio, can be designed to survive common forms of content processing. These represent one element of a response. However, the limitation of such approaches is the same as that which applies to all voluntary technical standards: their effectiveness depends on widespread adoption, and widespread adoption requires industry-level coordination and, potentially, regulatory mandate. The observation that governance in this domain cannot be accomplished by any single actor alone captures a structural feature of the provenance challenge that applies beyond consumer AI: it is a coordination problem as much as a technical one.

3.5. The end-of-life gap

Among the life cycle stages of an AI system deployed in defence and security contexts, the end-of-life stage has received, by some distance, the least attention, whether in academic literature, in policy discussion or in the governance frameworks currently under development. While increasingly acknowledged in policy documents, including at the level of the United Nations General Assembly,[17] the depth and breadth of reflections surrounding the end-of-life remain limited.[18]

An AI system that is retired, sold to a third party, repurposed for a different operational function or simply discontinued carries governance obligations that do not end at the point of procurement or deployment. The data that the system contains – including training data, operational logs, records of past engagement decisions – may remain sensitive long after the system itself is decommissioned. The capabilities it embodies may be reactivated or transferred. The forensic records it has generated may be relevant to accountability processes that occur after decommissioning. None of these dimensions are adequately addressed by governance frameworks that treat the life cycle as ending at the point when operational use stops.

The practical principle that emerged from the conference's discussions in this area is the same as that which governs the development and deployment stages: governance must be embedded in the architecture of the process, not retrofitted at its conclusion. A system that arrives at the end of its operational life without having been managed through a governed decommissioning process will leave unresolved accountability and data management questions that should have been addressed throughout its operation. The operators, commanders and policymakers who have relied on that system's capabilities are entitled to clear information about when those capabilities are no longer available and why; and the institutions responsible for managing accountability for past decisions have an interest in the integrity of the records those decisions generated.

17 General Assembly, "Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security", Report of the Secretary-General, A/80/78, 5 June 2025, https://docs.un.org/A/80/78.

18 Drew.

4. Data as technical foundation

4.1. Data as a fundamental prerequisite

Before AI systems can be designed, trained or deployed, a series of prior questions must be resolved: What data is available? Where did it come from? How has it been processed? And under what conditions can it be shared? These questions are not preliminary to the governance of AI; they are a central part of it. The technical infrastructure of AI – its performance, its reliability and its capacity to produce outputs that can withstand accountability scrutiny – rests on a data foundation whose quality, provenance and governance determine the value of everything built upon it.[19]

The relationship between data and AI has historically received considerably less attention in governance discussions than the AI systems themselves. Yet, as established in AISE25's sessions on data practices, this ordering inverts the actual logic of AI development: for all but the simplest AI applications, the vast majority of the effort required to make a system work well is spent on ensuring that the data on which the system is trained and evaluated meets the quality, accuracy and representativeness standards that the application requires.

The governance dimensions of data in the AI context can be organized around three questions: Where did the data come from and under what circumstances was it collected? How has it been processed, what other data has been linked with it and what laws or regulations apply to its use? And how can it be shared across organizational boundaries, with allied partners and for purposes of transparency or external scrutiny? Each of these dimensions creates specific governance requirements; governance failure at any one of them can undermine the reliability and accountability of the entire AI system built upon that data.

A further dimension concerns the relationship between data and the human beings who collect and interpret it. The characterization of data as neutral, as an objective record of reality, conflates the data itself with the decisions involved in its collection, labelling, organization and presentation. At every stage of the data life cycle, human choices shape what is captured, what is discarded and how what remains is categorized. Governance that attends only to the algorithmic dimensions of AI while ignoring the data dimensions will fail to address the most significant sources of bias, error and accountability risk.

4.2. Data governance as the 80 per cent problem: The example of the Dutch strategy

The most operationally grounded account of the challenges of data governance in a defence context at AISE25 came from the Netherlands' experience of developing and implementing its national defence AI strategy. The central finding of that exercise is that data governance accounts for roughly 80 per cent of the effort required to make an AI system work. Roles, responsibilities and operating principles for the governance of data are a prerequisite for any AI application. Where these foundations are not in place, the quality of data will determine outcomes regardless of the sophistication of the AI system built upon it.[20]

19 As noted by Calum Inverarity (Open Data Institute) in the facilitation of the data practices session on day 1.

20 van der Vlugt.

The governance architecture developed by the Netherlands in this context is organized around the insight that AI algorithms operate in specific domains (land, maritime, air, space, cyber) and that data-governance mechanisms must be aligned to those operational contexts, rather than imposed as a single enterprise-wide standard. This has led to a decentralized model in which commands within the defence structure carry their own data-governance responsibilities, operating within a central framework. The rationale is both organizational efficiency and operational accountability: those who use a system are best placed to govern the data on which it depends, provided that their governance practices are aligned with system-wide standards.

Two application cases illustrate the relationship between the quality of data governance and the performance of an AI system. A digital twin of naval vessels, developed in collaboration with maritime research and academic partners, required substantive decisions about which data to include and how to weight competing considerations before the AI model could be constructed; the governance of those decisions was itself the primary technical challenge. Separately, the development of an isolated, Internet-disconnected LLM environment for defence personnel, developed in response to the informal uptake of publicly available tools with sensitive information, required privacy and security requirements to be designed into the data architecture before model training could begin so that they would not bring subsequent constraints.

The question of digital sovereignty adds a further dimension to governance: where the critical infrastructure on which data pipelines depend is owned by actors in the commercial market, the governance leverage that a state has over its own AI systems is correspondingly constrained. This observation is not unique to the Netherlands, but it illustrates a structural feature of data governance for AI in defence contexts that is rarely addressed directly in governance frameworks: the state's ability to govern its own AI systems is, in many cases, limited by its dependence on commercially owned data infrastructure.

4.3. Data in conflict versus peacetime

The governance of AI data in defence and security contexts cannot be reduced to a single regulatory framework because the legal and ethical conditions under which data is collected, processed and used differ fundamentally depending on whether the context is armed conflict or peacetime. The same technological capability (e.g., drone sensor data) operates under different legal constraints, serves different functions and raises different accountability questions depending on context. Governance frameworks that do not distinguish between these contexts will either impose inappropriate constraints on legitimate peacetime applications or fail to provide adequate safeguards for data use in conflict.[21]

This observation was developed at AISE25 through a presentation of the data life cycle as comprising 11 stages – from initial generation and collection through curation, processing, storage, management, analysis, visualization, interpretation, use and eventual adjustment or deletion – at each of which specific validation criteria can be applied. Those criteria encompass not only technical dimensions such as accuracy, reliability and security, but legal and ethical ones including compliance with applicable classification requirements, military ethics and the avoidance of bias.

21 As noted by Kazuo Noguchi (Hitachi America, Ltd.) in a lightning talk on day 1: "AI and Human Data in Conflict vs Peacetime".

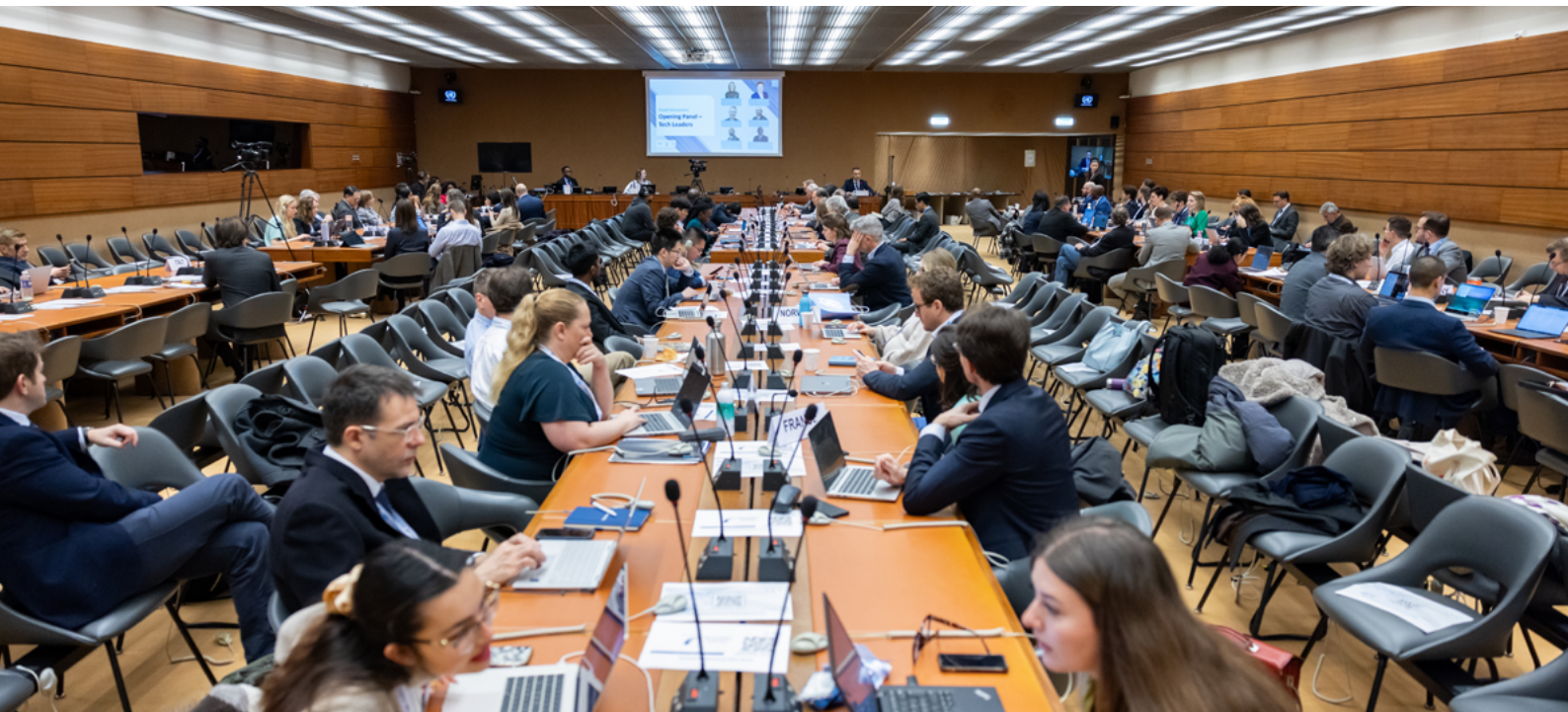
The governance implication is that data management for AI in defence contexts is a staged process at each stage of which specific decisions must be made and documented in a way that can withstand subsequent scrutiny. The question of how data practices in peacetime compare to and interact with those in conflict is one that governance frameworks have not yet addressed systematically, but that becomes more pressing as the integration of AI into security operations deepens. This includes the question of the purposes for which data about military personnel and civilian populations collected in peacetime may be used in conflict.

4.4. The provenance and watermarking frontier

As AI-generated content becomes increasingly indistinguishable from content produced by human actors, the provenance of both training data and AI outputs becomes a governance challenge. For the military and security domain, this challenge has a specific and immediate operational dimension: the attribution of intelligence products, targeting assessments and communications to their actual source, and the ability to distinguish AI-generated from human-produced content, is a prerequisite for the kind of accountability that responsible use of these technologies requires.

The governance problem is not merely about deception. As examined in Section 3.4, partial labelling strategies create exploitable vulnerabilities: in an adversarial environment where one party has strong incentives to manipulate the information environment of another, implicit trust signals about unlabelled content can be deliberately exploited. Governance frameworks that treat content labelling as an adequate response to the AI-generated content challenge will fail to address this structural dimension.

Technical approaches to content provenance represent one element of a response. These can include the embedding of imperceptible but machine-detectable watermarks in AI-generated images and audio, designed to survive common forms of content processing such as cropping and compression. However, the limitation of such approaches at scale is that their effectiveness depends on adoption across a sufficiently broad ecosystem of actors to make the absence of a watermark a meaningful signal. This is a coordination problem that requires multi-stakeholder engagement across industry, government and international standards bodies; it is a problem that no single developer or state actor can resolve unilaterally.



5. Trustworthiness, assurance and accountability

5.1. Responsible AI and assurance: The essential distinction

Among the cross-cutting governance challenges that AISE25 identified in the technical domain, the proliferation of terminology around AI ethics and governance deserves particular attention, not as a semantic concern but as a practical one with direct implications for accountability.[22]

The term "responsible AI", as it has been adopted by the defence community, reflects a meaningful adaptation from its civilian origins: it places less emphasis on values such as societal well-being, privacy and explainability, and more emphasis on reliability, safety and governability. These adaptations are, in themselves, defensible: they reflect the distinct legal, ethical and operational context of defence. They illustrate, however, a broader pattern in which the same term carries sufficiently different meanings across communities. This can mean that apparent agreement on vocabulary conceals substantive disagreement on requirements. If the term "responsible" means different things to civilian technology developers, defence actors and international security practitioners, then governance frameworks built around shared vocabulary do not necessarily build shared requirements.

Moreover, the term "responsible AI" could itself be a misnomer: it could be shorthand for responsible design, development, testing, evaluation, deployment, use and end-of-life cycle decommissioning of AI. When that chain of actors and stages is collapsed into a single adjective, the specific responsibilities at each stage – whether human, organizational or systemic – are obscured along with the architecture that responsible governance requires.

The specific relationship between responsible AI and AI assurance is a case in point. These are not synonymous terms:

- Responsible AI is a normative framework, an account of the requirements that an AI system should meet and of the obligations of the various human actors involved in its development and deployment.
- AI assurance is a process, a methodology for measuring, testing and communicating the degree to which an AI system meets specified requirements.

Each needs the other: responsible AI without assurance produces well-intentioned declarations that cannot be verified; AI assurance without responsible AI as its normative foundation can, in principle, measure a system against any assurance specification, including harmful ones. The governance implication is significant: responsible AI functions as the requirements-generator for AI assurance, and a governance architecture that conflates the two will systematically fail to close the gap between principles and accountability.[23]

The concept of trustworthiness thus offers a more tractable governance target than trust. Trust has hundreds of definitions and resists measurement; in contrast, a trustworthy system can be defined operationally as one that, when used correctly, reliably does what it is supposed to do and reliably does not do

22 As noted by Kerstin Vignard and Jane Pinelis (Johns Hopkins University Applied Physics Laboratory) in a dedicated deep-dive on day 1: "Responsible AI vs AI Assurance – A Semantic Showdown".

23 Vignard and Pinelis.

what it is not supposed to do. This formulation is context-specific: a system assured for one operational context requires a new assurance case if deployed in a different context; but it is measurable and amenable to the generation of evidence.[24]

A further concern raised at AISE25 was a trend, visible in recent policy and institutional developments, to move away from established safety-oriented terminologies in favour of new framings. The governance risk of this drift is not merely terminological: where governance frameworks and responsible AI communities have built shared understandings of what safety or ethical requirements mean in technical terms, the replacement of that vocabulary with new terms severs the connection between normative commitments and the technical communities that have developed methods for operationalizing them. The result is the proliferation of new adjectives (e.g., trustworthy, assured, safe, responsible) that create the illusion of governance progress while potentially eroding its substance.[25]

5.2. Digital forensics and the explainability prerequisite

Accountability for the use of AI in defence and security contexts – in the sense of a meaningful ability to reconstruct what happened, why it happened and who bears responsibility – depends on a technical prerequisite that many current AI deployments do not meet: explainability. Without the capacity to render a system's reasoning accessible to investigation, the tools of forensic accountability (i.e., examination of documentation, review of operational logs, reconstruction of decision sequences) cannot provide the evidentiary basis that legal and institutional accountability processes require.[26]

The application of digital forensics to AI systems deployed in military contexts involves, in principle, a chain of inquiry that begins not at the moment of an incident but at the point of procurement. The chain starts with the documentation associated with how a system was acquired and specified; the test results generated during development and validation; the record of how known problems were identified and addressed; the training data on which the system was built; and the source code, including any open-source or third-party components and the research papers on which the system's approach was based.

In practice, two structural features of the current landscape make this chain of inquiry difficult or impossible to complete. The first is classification: many AI systems deployed in military contexts are subject to security classifications that limit investigators' access to the documentation and source code that would otherwise enable accountability review. The second is complexity: modern AI systems, particularly LLMs with hundreds of billions of parameters, generate volumes of operational logs that are not practically reviewable without the assistance of other AI systems to process them. In the absence of explainable AI built into the system's design from the outset, forensic investigators face an accountability problem that is structurally similar to those arising in other areas of complex sociotechnical issues or even failures: the evidence exists in principle, but its volume and opacity make meaningful review functionally impossible without dedicated investment.

---

24 Ibid.

25 Ibid.

26 As noted by Drexx Laggui (Cybercrime Investigation and Coordinating Center, Philippines) in a dedicated deep-dive on day 1: "AI on the Battlefield – Building Trust and Accountability Through Digital Forensics".

The implication for governance is direct: explainability is perceived as not only a desirable technical feature, but as a precondition for accountability. Accountability gaps are created by design by governance frameworks that do not require explainability as a design specification for AI systems deployed in contexts where accountability may be required. As one formulation at the conference put it, the challenge is one of working on crimes that belong to the 21st century with 20th-century laws and 19th-century investigative tools.[27]

5.3. The skills gap: A technical governance problem

A governance challenge that does not receive adequate attention in discussions oriented primarily towards normative frameworks is the structural gap between the skills required to evaluate AI systems and those currently available in the organizations responsible for procuring, deploying and overseeing them. This gap is not a transitional problem to be solved as AI capabilities mature; it is defined by the relationship between AI development and AI governance.[28]

The core of the gap is that AI development has generated substantial training infrastructure and a large community of practitioners who are skilled in building AI systems. Meanwhile, the corresponding capacity to evaluate whether an AI system is reliable, to identify the conditions under which it will fail and to assess the adequacy of validation processes applied to it is far less developed. The capacity to make AI systems fail is categorically different from the capacity to build them, and it is rarely taught in technical training programmes. The result is that the organizations responsible for making governance decisions about AI systems (including about procurement, deployment and oversight) frequently lack the in-house expertise to evaluate the claims made by the developers of those systems.

An AI system that has been assessed as being reliable by its developer but which is deployed by an organization with limited capacity to verify that assessment has, in effect, no meaningful governance of its reliability. The development of independent evaluation and red-teaming capacity in defence institutions, regulatory bodies and oversight organizations is as much of a governance priority as the development of normative frameworks.[29]

A specific aspect of this challenge concerns the capacity to communicate technical knowledge in terms accessible to decision makers, and vice versa. While the primary addressees of international governance frameworks for AI in security are policymakers, military commanders and diplomats, they routinely receive technical, legal and ethical details in forms that they cannot necessarily operationalize (although they may be well understood by experts). The translation of complex technical and governance analysis into accessible language, without losing essential nuance, is itself a substantial contribution to governance, and one that experts bear a responsibility to make.[30]

The pace of AI development adds a further layer of complexity: AI systems are not static, and a system whose performance characteristics were established at one point in time may behave differently as the models underlying it are updated, as the data environment in which it operates changes and as the adversarial

27 Laggui.

28 As noted by David Sully (Advai) in day 2's opening panel on tech leaders.

29 As noted by Jibu Elias (Mozilla Foundation) in day 2's opening panel on tech leaders.

30 As noted by Pak Shun Ng (Ministry of Defence, Singapore) in day 2's panel on knowledge and capacity-building.

landscape evolves. The governance capacity required is accordingly not a one-time investment; it is an on-going commitment that must be streamlined across all layers of governance. Moreover, as AI deployment in security and defence contexts expands, the demands of this commitment will grow, rather than diminish.[31]

6. Conclusion: Carrying the conversation forward from AISE25 to AISE26

AISE25 served as a useful waypoint to take stock of the technical landscape for AI in defence and security as at March 2025 – not as a comprehensive survey of all systems and applications, but as a structured account of the technical realities, governance gaps and operationalization challenges that characterize the field at a specific moment in its development. AISE26 will convene at a moment when the normative and institutional frameworks examined in the accompanying report on governance will have advanced considerably, and when the question of what those frameworks require in practice – at the levels of system architecture, validation methodology and data management – will be more pressing than ever.

AISE26 will, in fact, meet after a dense period of institutional activities. Some of the key governance milestones that have been achieved between AISE25 and AISE26 include:

- The publication of the United Nations Secretary-General's report on AI in the military domain and its implications for international peace and security[32]
- The adoption by the United Nations General Assembly of resolution 80/58, a follow-up to resolution 79/239 on AI in the military domain and its implications for international peace and security,[33] which led to the organization of a three-day informal exchange on this issue to be held in Geneva on 15–17 June 2026[34]
- The organization of the third Responsible AI in the Military Domain (REAIM) Summit, held in A Coruña, Spain, in February 2026
- The publication of the Strategic Guidance Report of the Global Commission on Responsible AI in the Military Domain (GC REAIM)[35]
- The launch of the UNIDIR-led project for the development of a Framework of Responsible Industry Behaviour for AI in the Military Domain, to be developed in partnership with the Office of the United Nations High Commissioner for Human Rights (OHCHR)

---

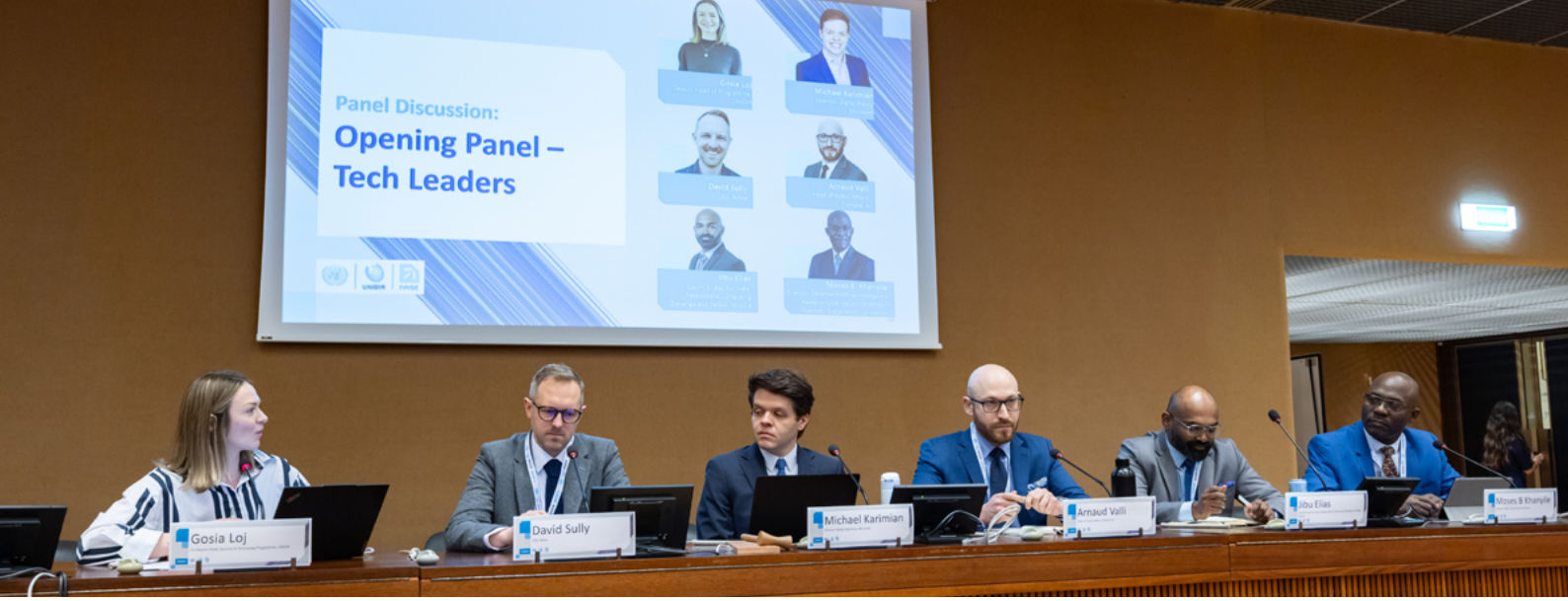
31 As noted by Moses B. Khanyile (Defence Artificial Intelligence Research Unit (DAIRU), Faculty of Military Science, Stellenbosch University, South Africa) in day 2's opening panel on tech leaders.

32 General Assembly, A/80/78.

33 General Assembly, resolution 80/58, "Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security", 1 December 2025, https://docs.un.org/A/RES/80/58.

34 Office for Disarmament Affairs, "UNODA Science, Technology and International Security Unit – Meeting", Meeting Place, 2026, https://meetings.unoda.org/unoda-stu-meeting/unoda-science-technology-and-international-security-unit-meeting-2026.

35 Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM), *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain* (The Hague: GC REAIM, September 2025), https://hcss.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf.



and in consultation with industry representatives and governments[36]

Amid these advances in governance, the technical dimensions of the AI governance challenge have not resolved themselves in the interim. The consistent finding across the technical sessions of AISE25 was that governance frameworks will not close the gap between principles and accountability unless they engage specifically with what those principles require at the level of system design, validation practice and data management. The translation of normative frameworks into technical requirements and the translation of technical knowledge into governance-accessible language remain the central operationalization challenges.

AISE26 is well-positioned to advance these translations. It can draw on the substantive technical expertise represented in its multi-stakeholder community and on the institutional experience accumulated by the processes and initiatives listed above. And it can do this in ways that produce specific, verifiable and actionable guidance for the practitioners who need it most.

As such, AISE26 is poised to advance some of the technical governance questions for which AISE25 paved the way, including, among others:

- What threshold of reliability is sufficient for the deployment of an AI system in operational defence contexts where failure carries significant consequences? What institutional infrastructure, including independent testing and evaluation capacity, is required to determine, verify and maintain that threshold?
- How should technical validation requirements (including the principles of auditability, controllability, transparency and security) be translated into procurement standards that can be applied consistently across different acquisition contexts? What mechanisms at the national and international levels could support consistent application of these technical validation requirements?
- At what point does the aggregation of individually validated AI systems into larger, operational architectures and systems create system-level risks that single-system validation can neither necessarily detect nor account for? How should governance frameworks account for this dimension of AI integration in defence and security contexts?

36 UNIDIR, “Framework of Responsible Industry Behaviour for AI in the Military Domain”, 2026, https://unidir.org/framework-of-responsible-industry-behaviour-for-ai-in-the-military-domain/.

Acknowledgments

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. This report was produced by UNIDIR's Security and Technology Programme, which is supported by the Governments of France, Germany, Italy, the Netherlands, the Republic of Korea and Switzerland, and Microsoft for its work on AI and autonomy. In addition, the Global Conference on AI, Security and Ethics 2025 was also supported by Advai.

UNIDIR extends its most sincere gratitude to all speakers, moderators, poster authors and audience members for their insightful presentations, comments and other contributions, which form the foundation of this report. UNIDIR also extends its profound gratitude to the members of the jury who helped review the abstracts received from UNIDIR's open call for proposals, which contributed to the conference's high calibre and the quality of the discussions: Dr. Alexi Drew, Major Jamal Mohamed Hassan, Calum Inverarity, Michael Karimian and Dr. Magdalena Pacholska. The evaluations were presented in the jury members' independent capacity and do not necessarily reflect the views or opinions of the jury members or of the organizations with which they work.

The authors wish to thank Dr. Giacomo Persi Paoli (UNIDIR) for advice, guidance, vision and support for the Global Conference and for his review of this report. Thanks for their support throughout the organization of the conference are also due to Jessica Lee Abowitz, Sapar Annayev, Asa Cusack, Jessica Espinosa Azcárraga, Anna Grangier, Edward Madziwa, Federico Mantellassi, Claudia Marquina and Mireia Mas Vivancos (UNIDIR), as well as to Elucidate Studios for the design support.

About UNIDIR

UNIDIR is a voluntarily funded, autonomous institute within the United Nations. As one of the few policy institutes worldwide that focus on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, it assists the international community in developing the practical, innovative ideas needed to address critical security problems.

About the UNIDIR Security and Technology Programme

Contemporary developments in science and technology present both new opportunities and challenges to international security and disarmament. UNIDIR's Security and Technology Programme aims to build knowledge and awareness about the international security implications and risks associated with specific technological innovations. It also convenes stakeholders to explore ideas and develop new approaches to address these issues.

Note

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries. The views expressed in the publication are the sole responsibility of the individual authors. They do not necessarily reflect the views or opinions of the United Nations, UNIDIR, its staff members or sponsors.

Citation

Yasmin Afina and Jan Hendrik Mannsperger, "AISE Markers Series – Benchmark II: Technology, Insights from the 2025 Global Conference on AI, Security and Ethics'", Geneva: UNIDIR, 2026.

About the authors

This report was produced by UNIDIR's Security and Technology Programme. It was drafted by Dr. Yasmin Afina and Jan Hendrik Mannsperger.

Photo credit

Cover photo generated with AI. Credit: Adobe Stock / Chaosamran_Studio. All other photos featured were taken by Pierre Albouy at the Global Conference on AI, Security and Ethics 2025, 27–28 March 2025, Geneva. Credit: UNIDIR / Pierre Albouy.

Acronyms and abbreviations

ACTS	Auditable, Controllable, Transparent and Secure (framework)
AI	Artificial intelligence
AISE	Global Conference on AI, Security and Ethics
LLM	Large language model
RAISE	Roundtable for AI, Security and Ethics

UNIDIR

@unidir
/unidir
/un_disarmresearch
/unidirgeneva
/unidir

Palais des Nations
1211 Geneva, Switzerland

UNIDIR.ORG