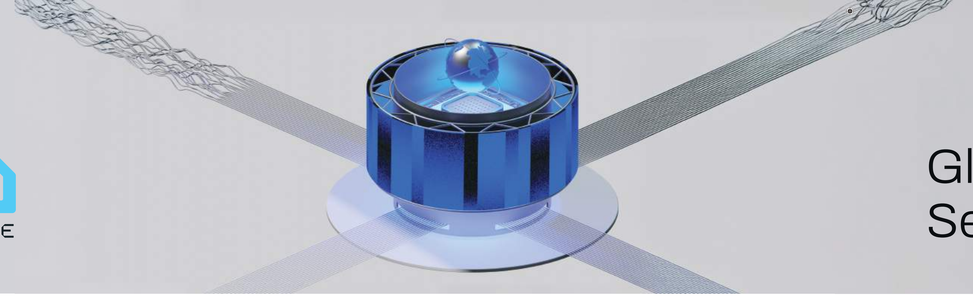


UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

unidir.org
#AISE26

ARTIFICIAL INTELLIGENCE DRIVEN DECISION MAKING ON MODERN WARFARE

Major Abel Murimi and Col (Dr) James Kimuyu Ph.D

JCSC Karen a College of National Defence University-Kenya

INTRODUCTION

Modern warfare has changed dramatically as a result of the incorporation of artificial intelligence (AI) into military operations, enabling faster, data-driven, and more precise decision-making processes. In Africa, AI incorporation is still at development phase, which has led to increased human errors when incorporating the AI algorithm, leading to insufficiency towards attaining its potential in combating crime. Addressing these issues requires targeted investment in ICT infrastructure, continuous capacity building of military personnel, and the development of clear regulatory and ethical frameworks to guide AI adoption.

OBJECTIVE

The objective of this study is to examine the effectiveness of artificial intelligence-driven decision-making in modern warfare within Africa, with a focus on how personnel competence, ICT infrastructure, and policy frameworks influence its adoption and performance in the Kenya Defence Forces.

METHODS

Cross-sectional research design was adopted since it allows the simultaneous collection of numerical data through surveys and descriptive insights through interviews, thereby enriching the interpretation of findings. Both quantitative and qualitative data were gathered using interview guides and questionnaires. SPSS version 27 was used to clean and code questionnaire data for descriptive statistics including mean, standard deviation, and frequency. Thematic approaches were the primary technique employed in the analysis of qualitative data in order to infer meaning from notes made during interview sessions.

RESULTS

The study found that AI has improved decision-making in modern warfare within the KDF by enhancing situational awareness, reducing human error, and strengthening threat prediction, although challenges such as inadequate training, limited ICT infrastructure in some areas, weak inter-departmental coordination, financial constraints, and slow policy updates still hinder full effectiveness.

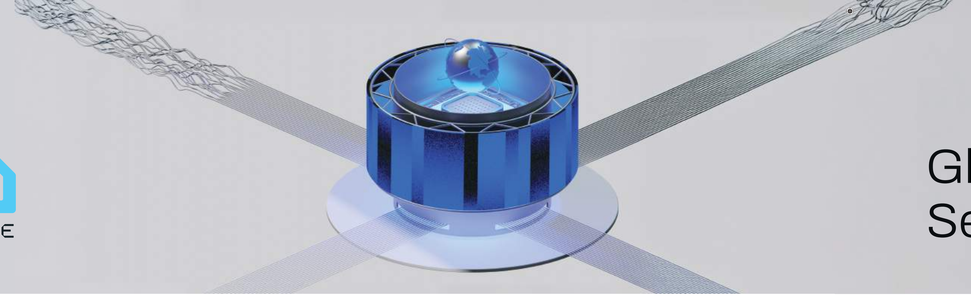
It concluded that effective AI-driven military operations depend on strong critical thinking, intelligence analysis, communication, and ethical judgment skills among personnel, supported by existing ICT infrastructure such as surveillance systems, communication networks, data centres, and cybersecurity frameworks, as well as policy instruments like the Kenya Defence Forces Act and Defence ICT Strategy, though these require continuous updating to remain relevant.

CONCLUSION

Effective AI-driven military operations depend on critical thinking, intelligence analysis, communication, and ethical judgment skills. Existing ICT infrastructure — including surveillance systems, communication networks, and cybersecurity frameworks — supports current AI operations, but gaps in rural connectivity and computing capacity persist. Formal policies provide accountability and ethical norms, but insufficient review frequency weakens their relevance against rapidly advancing technology.

RECOMMENDATIONS

1. Strengthen continuous, specialised AI decision-making training programmes emphasising situational awareness, critical thinking, and ethical judgment.
2. Expand and upgrade ICT infrastructure in rural and remote operational areas to ensure equitable AI access across all KDF formations.
3. Institute regular reviews and updates of AI-related policies to keep pace with technological advancements, and reinforce ethical guidelines ensuring accountability, transparency, and compliance with international humanitarian law.



Global Conference on AI, Security and Ethics 2026

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

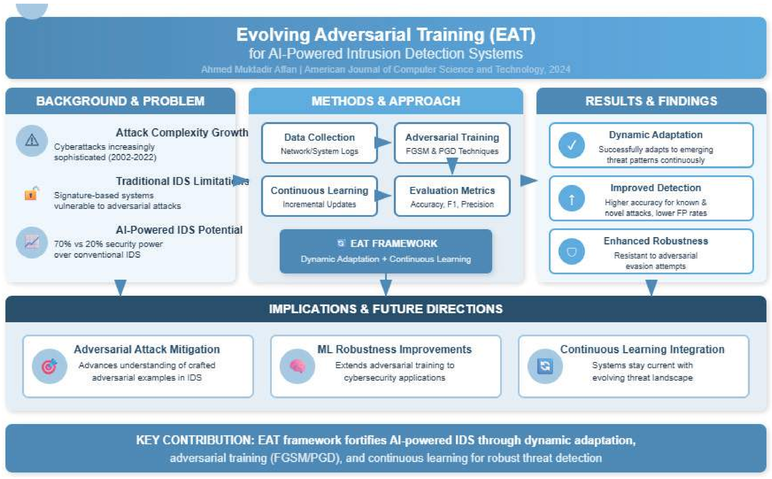
unidir.org
#AISE26

Evolving Adversarial Training (EAT) for AI-Powered Intrusion Detection Systems (IDS)

-Ahmed Muktadir Affan · ahmedmuktadir2007@gmail.com

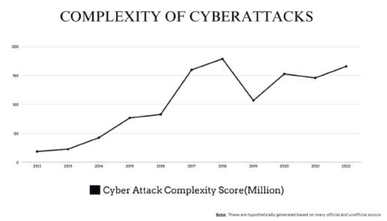
Introduction & Background

- Traditional signature-based intrusion detection systems (IDS) fail against evolving sophisticated cyberattacks due to their static models being vulnerable to adversarial attacks designed to deceive machine-learning algorithms.
- Cyberattack complexity has increased exponentially from 2002 to 2022, rendering traditional detection methods progressively less effective over time.
- AI-powered IDS demonstrate a higher security capability (70% vs 20% for conventional IDS), but integration of robust adversarial defence mechanisms is necessary.
- This study focuses on developing an adaptive, dynamic adversarial training (EAT) framework to enhance the robustness and effectiveness of AI-powered IDS.



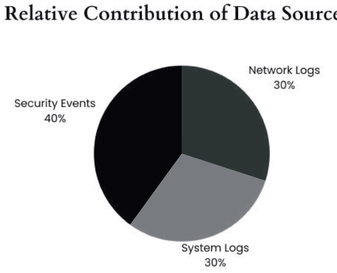
Methods

- Data collected includes historical intrusion datasets comprising network logs, system logs, and security events, with class imbalance addressed to improve dataset quality.
- The adversarial training pipeline introduces perturbed samples generated by the FGSM and PGD methods during model training to simulate realistic attack variations.
- Continuous learning mechanism incrementally updates the model with new data, retraining on recent attack instances to ensure adaptation to emerging threats.
- Performance evaluation includes metrics such as accuracy, precision, recall, and F1-score, emphasising the detection of novel attack patterns.

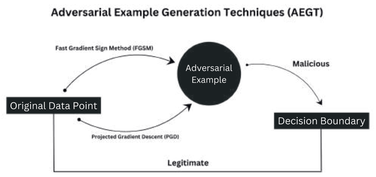


Cyberattack complexity score trajectory showing an increasing trend over two decades.

Security power comparison between conventional and AI-powered intrusion detection systems.



Data source composition for intrusion detection.



Adversarial Example Generation Techniques (AEGT)

Key Findings

- Evolving Adversarial Training (EAT) significantly enhances AI-powered IDS by enabling dynamic adaptation to evolving cyber threats.
- AI-powered IDS with adversarial training demonstrated a marked improvement in security power (70%) compared to conventional methods (20%).
- Continuous learning and adversarial example integration lead to increased detection accuracy and robustness against adversarial attacks.
- EAT framework holds promise for scalable integration into broader cybersecurity defence strategies.

Results

Dynamic Adaptation

- The IDS successfully adapts to changing threats by continuously learning and integrating new adversarial information from emerging attack patterns.
- Adversarial training using FGSM and PGD techniques enables iterative model exposure to varied attack perturbations.

Threat Detection Improvement

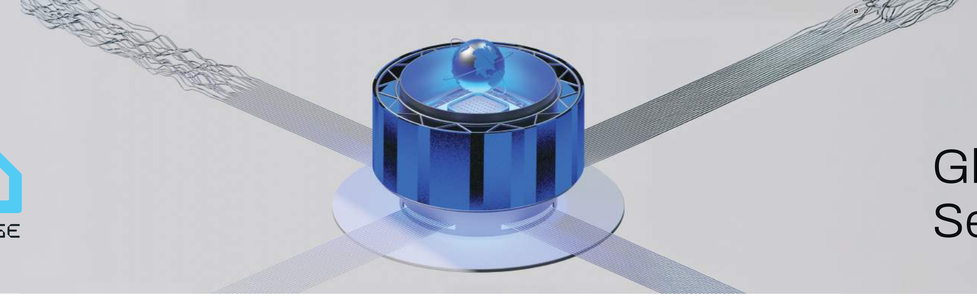
- Adversarial training exposure increases detection accuracy for both known and novel attacks while significantly reducing false positive rates.
- Incremental learning retrains models on recent attack data to improve the identification of emerging threat patterns.

Robustness and Generalisation

- EAT framework enhances model robustness against adversarial evasion attempts, making IDS resistant and generalizable across diverse threat scenarios.
- Reduced deceivability correlates with improved detection accuracy and overall generalization.

Discussion & Implications

- The findings confirm that dynamic adaptation via adversarial training improves the robustness of machine learning models, particularly in the cybersecurity domain.
- EAT advances the field by applying continuous learning principles to IDS, ensuring responsiveness to the evolving threat landscape.
- The approach reinforces defences against crafted adversarial examples that traditionally bypass static IDS models.
- Scalability and integration into broader cybersecurity systems suggest tangible applications for improved organisational security postures.



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

unidir.org
#AISE26

Behavioural Validity: A Missing Criterion in AI Governance for Security Applications

Dr Hina Tahseen, MBBS • MSc (Cardiff) • PGCert (Oxford) • MRCPsych
Consultant Psychiatrist & Responsible Clinician, Somerset NHS Foundation Trust • Honorary Lecturer, School of Medicine, Cardiff University •
Vice Chair, RCPsych Rehabilitation & Social Psychiatry Faculty • ORCID 0009-0000-3834-2864 • LinkedIn /in/drhinatahseen

ABSTRACT

AI systems deployed in security, border control and communications monitoring rely on behavioural signals to classify risk and intent. These systems are built on training data that overwhelmingly reflects normative behavioural patterns, systematically excluding presentations associated with severe mental illness, neurodivergence and psychological trauma.

The result is a governance gap. A person experiencing psychosis, mania or post-traumatic dissociation produces behavioural signals that AI may classify as threatening, deceptive or suspicious. Pressured speech registers as agitation. Disorganised thought is read as evasion. Fixed beliefs are flagged as potential radicalisation. These are clinical misreadings with real-world governance consequences.

This poster proposes behavioural validity as a necessary criterion in AI governance for security applications: the requirement that systems demonstrate reliable interpretation of human behaviour across the full clinical spectrum, not only normative patterns. It draws on clinical psychiatric experience to illustrate specific failure modes and argues that without clinical input into AI evaluation, governance frameworks will continue to assess technical performance while missing the human behaviours that matter most.

FIVE CLINICAL FAILURE MODES

	Presentation	What AI infers
1	Pressured speech, rapid topic shift, reduced sleep need [MANIA]	Agitation · hostility · intoxication
2	Disorganised thought, tangential speech, loosened associations [PSYCHOSIS]	Evasion · deception · cognitive impairment
3	Affective incongruence: flat narration of traumatic content [TRAUMA · DISSOCIATION]	Deception · fabrication · rehearsed account
4	Atypical eye contact, prosody, response latency [AUTISTIC COMMUNICATION]	Suspicion · non-cooperation · concealment
5	Avoidance & distress around ego-dystonic intrusive imagery [OCD]	Preoccupation with violent or harmful intent

HISTORICAL CONTINUITY

2013 – TSA SPOT Programme: US DoD funds low-ridership screening at domestic hubs; deception detection: meta-analysis of 400+ studies shows accuracy ~66–80%.

2019 – BorderControl: Horizon 2020 €8M project; 24 European sites. Academic in Hungary, Greece, Latvia; scientific validity questioned in peer-reviewed literature (EU H2020 · LIBE · ITN).

2024 – EU AI Act: Regulatory recognition: prohibits deception systems in workplaces and education. Security and border applications remain high risk without AI (Art. 5(1)(f), Annex III).

Behaviourally invalid indicators embedded in AI → systematising the error at scale.

> *"Security AI may classify behaviour, but still misread distress, trauma or neurodivergent communication."*

THE MISREADING MECHANISM

OBSERVED BEHAVIOURAL SIGNAL:
Fragmented narrative · flat affect · avoidance · atypical eye contact

AI INFERS
Deception · Evasion · Concealment · Threat

CLINICAL REALITY
Trauma response · Dissociation · Autistic processing · Ego-dystonic distress

Same form. Different function. The system cannot tell.

THE GOVERNANCE GAP

The EU AI Act regulates emotion and behavioural recognition unevenly by deployment context. Existing frameworks audit technical accuracy, robustness, demographic fairness – not the behavioural taxonomy on which classification rests.

Prohibited: Workplaces · Schools · Universities

High-risk but allowed:
Borders · Asylum interviews · Monitored communications · Detention & security screening

The populations with the highest clinical and trauma burden remain exposed precisely where governance is weakest.

PROPOSED CRITERION

***Behavioural validity** — the requirement that AI systems demonstrate reliable interpretation of human behavioural signals across the full clinical and neurodivergent spectrum, not only normative patterns.*

WHO CARRIES THE COST

Populations with the highest exposure to AI-enabled behavioural classification carry with them the highest prevalence of psychiatric morbidity, neurodevelopmental difference, and trauma-related presentations:

- Displaced persons and asylum applicants
- Conflict-affected and post-conflict communities
- People in detention or under monitored communications
- Survivors of torture and complex trauma

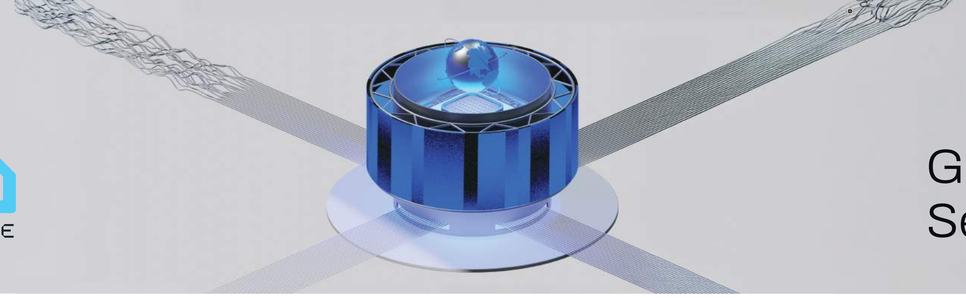
> *"AI flags affective incongruence as deception. Psychiatry knows it as trauma, dissociation, masking, depression."*

AI governance for security applications should incorporate behavioural validity alongside technical performance. Without clinical input, governance will continue to assess what AI does, while missing whom it fails.

KEY REFERENCES

1. Barrett et al. (2019). Emotional expressions reconsidered. *Psychol Sci Public Interest*.
2. Herlihy et al. (2020). Discrepancies in autobiographical memories of asylum seekers. *BMJ*.
3. Whittaker et al. (2019). Disability, Bias and AI. AI Now Institute.
4. US GAO (2013). GAO-14-159: TSA behavior detection Initiations.
5. Regulation (EU) 2024/1689 — AI Act. Art. 5(1)(f), Annex III.

READ FURTHER •
drhinatahseen.substack.com



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

unidir.org
#AISE26

Implications of Adversarial AI on strategic stability

Adversarial AI refers to the manipulation of data to disrupt access to AI systems, generate misleading outputs, or expose sensitive information. As AI is increasingly integrated into civil and defence applications, the attack surface for adversarial AI is expanding in scale and complexity.

These fast-evolving adversarial AI threats should not slip through the seams of broader AI governance initiatives. Although AI is still being developed and integrated into military systems, meaning adversarial AI may only threaten stability in the longer term, there is a real and current opportunity to proactively strengthen systems and international governance.

MECHANISM OF TWO TYPES OF ADVERSARIAL AI ATTACKS

PATHWAYS TO STRATEGIC STABILITY IMPLICATIONS

Pathway	Mechanism
Undermined deterrence	Perceived AI vulnerability erodes credibility
Offence-defence balance	Easier to attack AI than defend it, favouring aggressors, potentially eroding trust and leading to hyper-caution
Ambiguity	Difficulty distinguishing attacks from malfunctions creates escalation risk
Empowered non-state actors	Democratised adversarial AI techniques complicate attribution

KEY FINDINGS

- Every military AI capability is simultaneously a new attack surface for adversarial manipulation.
- Current governance focuses on technical solutions (TEVV, robustness) but underexplores resilience measures and escalation containment.
- Adversarial AI falls between existing governance silos (cyber, nuclear, conventional arms), risking gaps.

RECOMMENDATIONS

- **Embed adversarial AI in existing governance**
- **Develop international norms** around adversarial AI and what constitutes responsible behaviour in the deployment, compromise or targeting of different AI systems in a defence context.
- **Promote after-action reviews** to assess the source, impact, and systemic implications of attacks, alongside a commitment to taking steps to prevent recurrence
- **Promote system resilience** in governance (i.e., ensuring rapid restoration of function).

METHODS

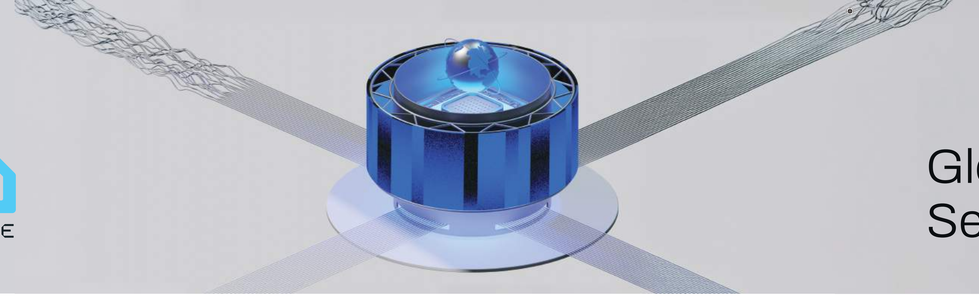
This mixed-method qualitative study (September 2025–February 2026) combined a Rapid Evidence Assessment of academic and grey literature, an expert workshop with AI governance and strategic stability specialists, key informant interviews across government, academia and industry, scenario analysis of five illustrative attacks, and a gap analysis of existing governance mechanisms.

CONTACT

Dr. Irene van Droffelaar, Senior Research Data Scientist at RAND Europe
irene@randeurope.org

This study was conducted by RAND Europe, and supported by the UK Foreign, Commonwealth & Development Office

RAND EUROPE | Foreign, Commonwealth & Development Office



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

unidir.org
#AISE26

ALGORITHMIC R2P
AI, Gendered Disinformation, and the Duty to Protect in Conflict

BACKGROUND

AI-mediated disinformation, amplified by algorithmic bias and narrative manipulation, can inflame societal divisions and act as a precursor to real-world violence.

This research proposes *Algorithmic R2P:* a framework that extends the international Responsibility to Protect (R2P) to the digital sphere, establishing a duty to protect populations from AI-mediated information harms that can trigger or exacerbate conflict.

IMPLICATIONS & RECOMMENDATIONS

- Integrate AI governance with conflict prevention frameworks through an Algorithmic R2P lens.
- Mandate algorithmic impact assessments and narrative risk audits for high-risk AI systems.
- Invest in OSINT integrity tools (e.g., C2PA, provenance tech) and capacity building for the Global South.
- Foster multi-stakeholder cooperation for rapid information integrity response mechanisms.

THE ALGORITHMIC R2P FRAMEWORK

Extending the Responsibility to Protect to the Digital Information Environment

1	2	3	4	5
ANTICIPATE RISK	**PREVENT HARM**	**PROTECT PEOPLE**	**RESPOND & COOPERATE**	**ENSURE ACCOUNTABILITY**
Identify and assess AI systems that can amplify harmful narratives or disinformation.	Mitigate risks through algorithmic impact assessments, robust content verification, and bias testing.	Safeguard populations from targeted disinformation and narrative manipulation that can incite violence.	Enable rapid information integrity response and cross-border cooperation when harms emerge.	Establish traceability, attribution, and redress mechanisms for actors and systems.

RESEARCH OBJECTIVES

1. Conceptualise "Algorithmic R2P" as a governance architecture linking human rights, security, and AI ethics.
2. Identify how AI-generated and algorithmically amplified content can trigger communal tensions and violence.
3. Propose a multi-stakeholder framework for proactive mitigation and accountability.
4. Offer actionable recommendations for policymakers, platforms, and practitioners.

GUIDING PRINCIPLES

Human Dignity | Do No Harm | Transparency & Explainability | Equity & Inclusion | Proportionality & Necessity | Peace & Security

HOW ALGORITHMIC R2P WORKS (RISK PATHWAY)

AI CONTENT GENERATION
Synthetic or manipulated content is created at scale (text, image, video, audio).

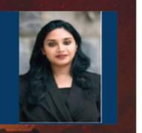
ALGORITHMIC AMPLIFICATION
Recommender systems prioritize sensational or polarizing content, amplifying it to vulnerable audiences.

NARRATIVE MANIPULATION
Demographic-targeted disinformation distorts identity, fuels stereotypes, and erodes trust.

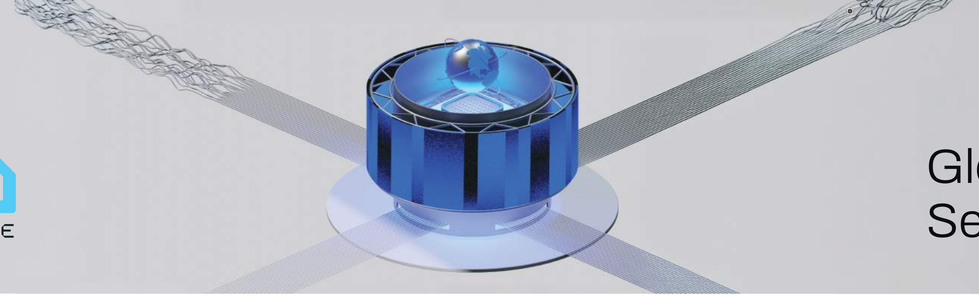
SOCIAL INFLECTION POINT
Online tensions spill over into offline mobilisation, protests, hate crimes, or communal violence.

CONFLICT ESCALATION
Information-driven polarisation contributes to broader instability and security threats.

Research Author:
Jeethu Cherian



If algorithms can amplify gendered harm in conflict, the Responsibility to Protect must evolve into an Algorithmic R2P. If conflict is increasingly shaped by algorithms, then the duty to protect must extend into the information domain.



Global Conference on AI, Security and Ethics 2026

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

unidir.org
#AISE26

GOVERNING THE UNGOVERNED: A FOUR LAYER FRAMEWORK FOR AI IN CONFLICT ANALYSIS AND SECURITY IN AFRICA

INTRODUCTION

Artificial intelligence is being deployed across conflict-affected environments in Africa at an unprecedented pace in early warning systems, humanitarian logistics, military operations, security surveillance and peacebuilding processes. Yet the governance architecture to make these deployments safe, accountable and legitimate does not exist at the scale or depth that the moment demands.

The central failure is structural. Policymakers and technologists ask whether AI systems are accurate, whether they predict correctly, whether they perform at scale. These are the wrong first questions. The right questions are: who governs this system, through what institutions, with what community consent, and what happens when it fails?

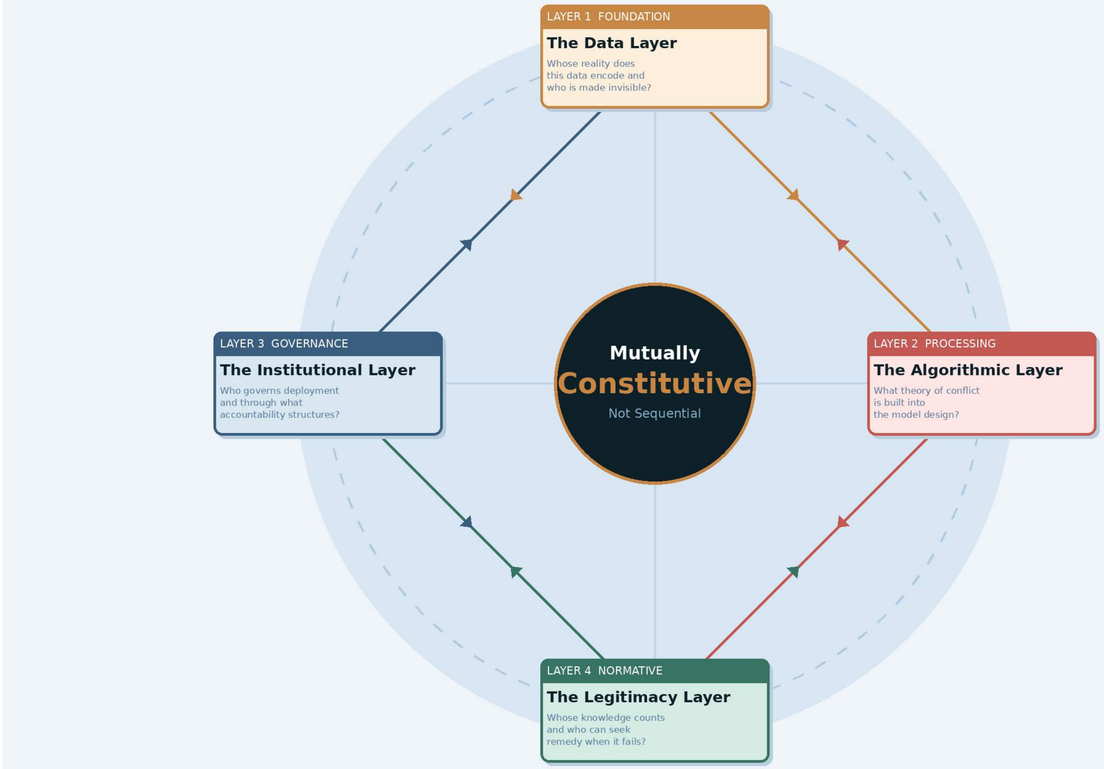
THE GOVERNANCE GAP

Current AI governance frameworks including the EU AI Act, national AI strategies, and emerging multilateral instruments address the technical and regulatory dimensions of AI deployment. They do not adequately address the institutional accountability structures or the legitimacy frameworks that determine whether AI systems are trusted, owned and contestable by the communities whose conflicts they are meant to address. General Assembly resolutions 79/239 and 80/58 mark a historic shift toward implementation of AI governance in the military and security domain. Yet even these landmark instruments focus predominantly on state behaviour and technical standards. The governance gap at the institutional and legitimacy layers remains. In Africa, where AI is being deployed in some of the world's most complex conflict environments, this gap is most acute and most dangerous.

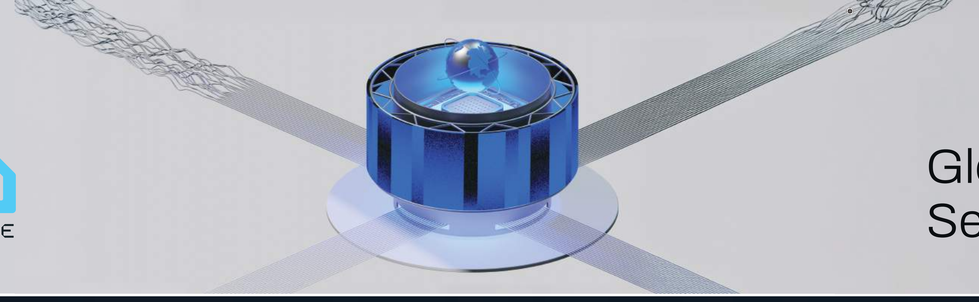
WHY AFRICA MATTERS

Africa is not a passive recipient of AI governance frameworks designed elsewhere. The AU Peace and Security Council has issued landmark decisions on AI and peace and security between 2024 and 2025. The November 2025 Kigali Roadmap adopted the first continental framework for AI integration into the Continental Early Warning System. The AU AI Advisory Group convened in Nairobi in December 2025. Africa is building its governance architecture right now at precisely the moment when its design can still be shaped by African voices, African research, and African institutional realities.

This poster contributes a governance diagnostic tool developed from that African institutional context.



Prudence Chepngeno | AI-Assisted Conflict Research, Security Analysis and Digital Peacebuilding | Nairobi, Kenya



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

unidir.org
#AISE26

INTEGRATING ARTIFICIAL INTELLIGENCE INTO CYBER DEFENCE OPERATIONS:

ENHANCING SOC WORKFLOWS AND INCIDENT RESPONSE

Using AI to empower analysts, accelerate decisions, and strengthen cyber resilience.

ABOUT ME

ZOE DUNCAN

Cybersecurity professional from Jamaica with experience in SOC operations, threat monitoring, OSINT, and cyber incident analysis. Passionate about building resilient cyber defence through AI, collaboration, and continuous learning.

LET'S CONNECT

linkedin.com/in/zoe-duncan-67ba26215

zoeduncan46@gmail.com

1 THE CHALLENGE

Security Operations Centres (SOCs) are overwhelmed by high volumes of alerts, complex threats, and ever-evolving attack techniques.

THE RESULT:
alert fatigue, delayed detection, and slower incident response.

2 THE GAP

- Current tools generate too many false positives and lack context.
- Analysts spend more time filtering noise than responding.
- AI is underutilized or deployed in silos, not aligned with SOC workflows.

3 OUR APPROACH

Embed AI into existing SOC workflows to augment human analysts, improve decision-making, and accelerate incident response.

AI acts as a force multiplier, not a replacement.

4. AI-ENHANCED SOC WORKFLOW

1 INGEST Data & Alerts
Collect logs, alerts, and telemetry from multiple sources into the SOC.

2 ANALYSE with AI
AI models detect patterns, anomalies, and enrich alerts with context and threat intelligence.

3 PRIORITISE Smart Triage
Alerts are scored and prioritised by risk and impact so analysts focus on what matters most.

4 INVESTIGATE Analyst Focus
Analysts investigate high-confidence, high-impact alerts with AI-provided insights and recommendations.

5 RESPOND & IMPROVE
Faster responses with automation (SOAR) where appropriate. Feedback improves AI models over time.

1000+ Daily Alerts Processed

20–50 High-Priority Alerts Escalated

FASTER RESPONSE Stronger Outcomes

SMALL ISLAND DEVELOPING STATES CONTEXT

- Resource-constrained settings
- Evolving threat landscape
- Building regional cyber resilience

5. KEY BENEFITS

- Faster detection and response
- Reduced alert fatigue
- Better prioritisation and accuracy
- Stronger, data-driven decision making
- Continuous learning and improvement

6. REAL-WORLD IMPACT

- Shorter mean time to detect (MTTD) and respond (MTTR)
- More efficient use of analyst time and expertise
- Improved resilience against evolving threats
- Scalable across organisations and regional partnerships

HOW IT HITS THE SOLUTION THEMES

OPERATIONAL RELEVANCE
AI supports detection, triage, and response in existing SOC workflows.

PRACTICAL IMPLEMENTATION
Tools and processes that are realistic for resource-constrained settings.

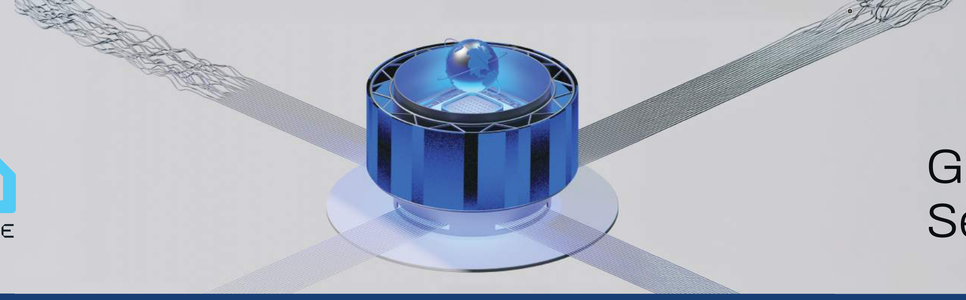
IMPACT
Delivers faster prioritisation and smarter decisions for Small Island Developing States.

LOOKING FORWARD
Future adoption can expand through regional collaboration and shared learning.

7. EXAMPLE USE CASES

- AI-driven anomaly detection in network traffic
- Phishing and malware detection using behavioural analytics
- Automated alert enrichment with threat intelligence
- AI-assisted incident investigation and reporting
- Threat hunting and proactive defence

AI + HUMAN EXPERTISE = STRONGER CYBER DEFENCE



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

unidir.org
#AISE26

Audit Before Deploy: *A Simple AI Risk Checklist for Governments*

BACKGROUND & RATIONALE

AI systems are increasingly deployed in security-sensitive government contexts-border control, public surveillance, and critical infrastructure. High-level governance principles exist but practical tools for rapid, non-technical assessment remain scarce.

The 'governance gap' between policy intent and operational practice is widest precisely where AI risks are highest. This paper addresses that gap with a structured, five-dimension pre-deployment checklist.

FIVE ASSESSMENT DIMENSIONS

1 Transparency & Explainability

Documentation completeness • Output explainability • Third-party auditability • Data provenance

2 Bias, Fairness & Non-Discrimination

Demographic performance testing • Data representativeness • Legal compliance • Ongoing monitoring

3 Dual-Use & Mission Boundaries

Mission scope definition • Repurposing potential • Data-sharing controls • Int'l law compliance

4 Human Oversight & Control

Human-in-the-loop • Override mechanisms • Operator competency • Accountability assignment

SCENARIOS & FINDINGS

Border Security

AI risk scoring raises transparency and bias concerns. Checklist surfaces gaps in demographic testing and dual-use potential of traveler data.

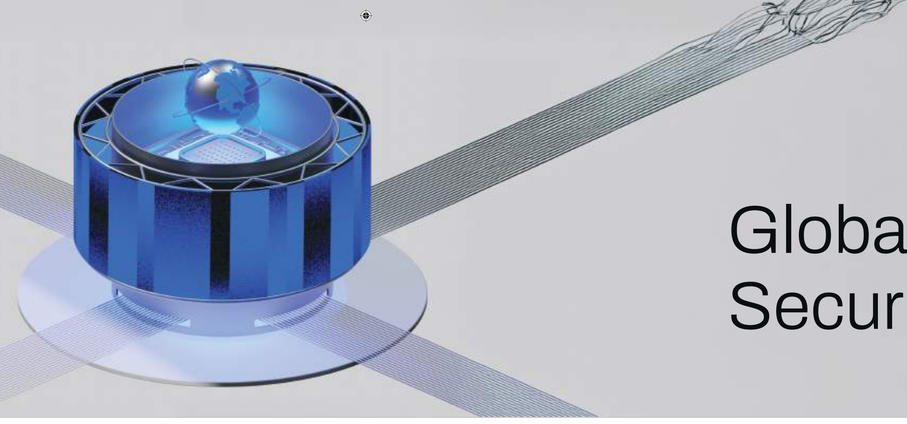
Public Surveillance

City-scale systems amplify even small false-positive rates. Failure mode and rights-impact criteria are critical.

Critical Infrastructure

Adversarial robustness and meaningful human control under operational time pressure are governance priorities.

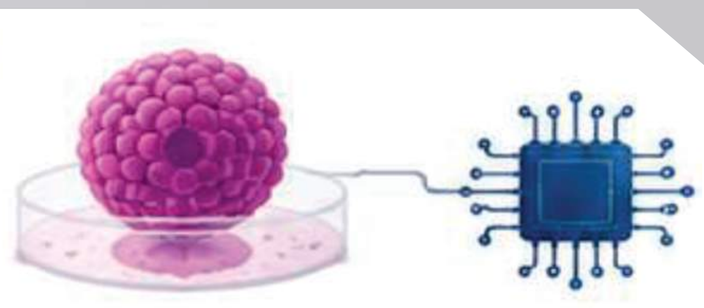
Kamal Tasiu
kmlts256@gmail.com



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

Post-Silicon AI and Global Security: Governing Organoid Intelligence



Prof. Dr. A. Inci Sökmen ALACA
Global Defence Information Center
incisokmen@gmail.com

RESEARCH QUESTION

How can the international community develop an inclusive, effective and anticipatory governance framework to manage the security, ethical and dual-use risks of organoid intelligence systems while enabling responsible innovation?

WHY IMPORTANT FOR SECURITY?

- Organoid AI technologies could transform the technological substrate of AI, creating new strategic advantages and risks.
- Dual-use potential may enable advanced decision-making, autonomous systems and strategic simulations with unpredictable behaviours.
- Bio-digital nature introduces novel biosecurity, ethical and control challenges.
- Fragmented governance and regulatory gaps may lead to misuse, arms race dynamics and regional or global instability.
- Anticipatory governance is essential to safeguard international peace, security and human values.

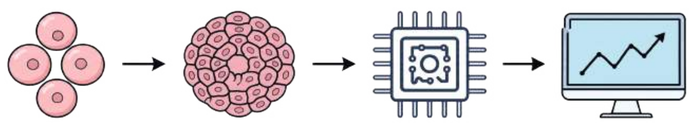
BACKGROUND

Recent advances in neuroscience, synthetic biology and artificial intelligence are enabling a new era of computation based on living neural tissue — organoid intelligence (OI).
These bio-computing systems represent a potential post-silicon paradigm that combines biological neurons with computational interfaces capable of adaptive learning and complex signal processing.

While OI holds promise for medicine, scientific research and energy-efficient computing, it also introduces unprecedented challenges for global security and technology governance.

WHAT IS ORGANOID INTELLIGENCE?

Organoid intelligence uses laboratory-grown clusters of neurons, called organoids, connected to computational systems.
These hybrid bio-digital systems can receive inputs, process signals, and exhibit learning-like behaviors.



Key features: biological adaptability, low energy potential, parallel processing, and emergent dynamics that differ from traditional silicon-based AI.

WHY IT MATTERS FOR GLOBAL SECURITY

AI technology is advancing faster than legislation can keep up, creating regulatory gaps that pose serious ethical and strategic risks.
Organoid AI systems could be leveraged in security-relevant applications such as decision support, autonomous platforms and strategic simulations, raising dual-use and proliferation concerns.
The biological substrate of OI introduces uncertainties related to predictability, control and ethical oversight.
Uncoordinated development may lead to governance gaps, potential misuse and regional instability.

A structured, responsible and inclusive governance approach is essential to ensure responsible innovation and international security.

SECURITY IMPLICATIONS

MILITARY DECISION SUPPORT
Potential use in strategic analysis, battlefield simulations and autonomous systems.

BIOSECURITY RISKS
Dual-use potential and risks of misuse or unauthorized access to biological systems.

ETHICAL AND NEUROETHICAL CONCERNS
Issues related to consciousness, suffering, identity and moral status of synthetic neural systems.

STRATEGIC COMPETITION
Emerging technology race may deepen inequalities and destabilize strategic balances.

UNCERTAINTY AND CONTROL
Unpredictable emergent behaviors and challenges in verification, validation and long-term control.

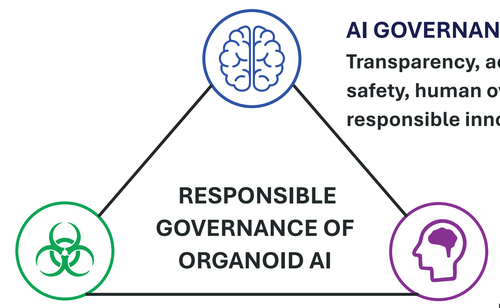
RISK MATRIX

Risk Category	Impact	Likelihood
Dual-use / Proliferation	High	High
Biosecurity Threats	High	Medium
Ethical / Neuroethical	High	Medium-High
Uncertainty / Control	Medium-High	High
Strategic Competition	High	High

Without governance, risks may outpace benefits and create long-term security vulnerabilities.

PROPOSED GOVERNANCE FRAMEWORK

An interdisciplinary framework integrating three key domains is essential for responsible governance of organoid intelligence.



Effective governance requires coordination among states, international organizations, academia, industry and civil society.

POLICY RECOMMENDATIONS

1 Establish International Monitoring Mechanisms
Create an international monitoring system to track advances in organoid AI, share information on bio, computing capabilities, and promote transparency.

2 Develop Biosecurity Standards
Develop biosecurity standards and risk assessment frameworks for organoid AI systems, including guidelines for safe research and laboratory practices.

3 Ensure Transparency and Auditing
Ensure transparency, auditability and bias mitigation through inclusive governance mechanisms and independent oversight.

4 Promote Multistakeholder Engagement
Engage governments, researchers, industry and civil society in developing ethical norms and governance principles.

5 Advance International Cooperation
Strengthen international cooperation to prevent dual-use misuse and ensure responsible, equitable and peaceful innovation.

KEY TAKEAWAYS

- Organoid intelligence represents a potential post-silicon shift in AI.
- Its bio-digital nature introduces new security, ethical and governance challenges.
- An interdisciplinary, anticipatory governance framework is essential.
- International cooperation and inclusive norms can ensure responsible innovation and global security.

LOOKING FORWARD

The governance of organoid intelligence will shape the future of artificial intelligence and international security. Through proactive policies, shared norms and scientific responsibility, the global community can harness the benefits of this emerging technology while minimizing its risks to humanity and peace.

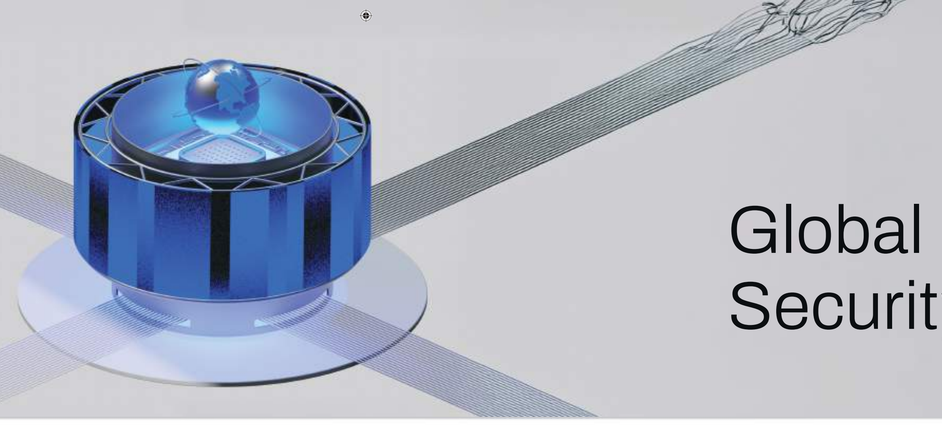
ABOUT ME

Prof. Dr. A. Inci Sökmen ALACA
Global Defence Information Center
incisokmen@gmail.com
Research Focus: AI Governance, Biosecurity, Neuroethics, Emerging Technologies and International Security

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



Global Conference on AI, Security and Ethics 2026

From Traditional to Agentic AI: Securing Autonomously Acting Language Models against Emergent Threats

Alyssa Columbus Johns Hopkins Bloomberg School of Public Health & Ludwig Maximilian University of Munich

Motivation: A New Threat Surface

Modern LLMs no longer behave as one-shot text generators. Embedded in ReAct-style loops, multi-agent systems, and tool-using assistants [3, 7], they autonomously plan, invoke APIs, modify files, and pursue objectives across long horizons. Production agents already book calendars, execute code, and broker transactions on behalf of users. The attack surface has shifted in lockstep:

- **Prompt injection** propagates through tool outputs, retrieval results, and inter-agent messages, often executing several turns after the original injection [1].
- **Privacy violations** extend beyond memorized training data [6]: an agent can exfiltrate policy, credentials, or contextual secrets through legitimate-looking tool calls.
- **Hallucinations become actions**: a hallucinated function argument now mutates state in external systems, with no opportunity for downstream correction.
- **Compositional risk**: behavior of a multi-agent system is not a sum of its parts; emergent failure modes appear only after several agents interact [4].

Perimeter-based defenses assume a clear input–output boundary that **agentic systems violate**. **Central claim:** securing agentic LLMs requires treating the *agent's behavior over time* as the primary object of defense.

Why Governance Audiences Should Care

AI assurance today (model cards, red-team reports) describes systems at a *point in time*. Agentic deployments need **life-cycle-wide telemetry** because risk emerges through the action trajectory, not at any single decision point [2, 8].

- **Assurance regimes** should require auditable action graphs, not only eval snapshots; vendors should disclose how telemetry is generated and retained.
- **Defense and security procurement** must ask how oversight policies compose across multi-agent systems and how decision authority is allocated between operators and autonomous components.
- **Confidence-building measures** between states can be grounded in shared telemetry schemas without exposing model internals or training data.
- **Dual-use concerns** sharpen as the same agentic capabilities support legitimate operations (logistics, intelligence analysis) and abuse (cyber operations, disinformation) [10].
- **International law** applies most cleanly when attribution chains are reconstructible from behavioral logs relevant to IHL distinction, proportionality, and post-incident review [10].
- **Standard-setting bodies** (ISO, NIST, OECD) can converge on common telemetry schemas for TEVV of agentic systems [2].
- **Liability frameworks** require traceability from agent action back to design choices and operator authorization; behavioral telemetry provides the audit trail.

From Perimeter Defense to Behavioral Monitoring

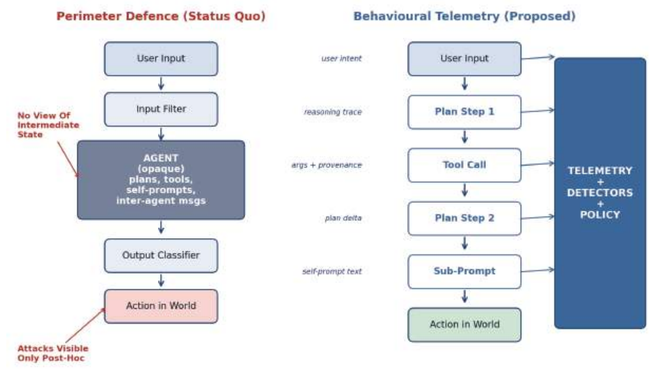


Figure 1. *Perimeter defenses treat the agent as opaque (left). Behavioral telemetry (right) treats every plan step, tool call, and sub-prompt as a first-class observable. Italic labels show what each event captures.*

Threat Taxonomy: Classic ML Failure Modes Mapped to Their Agentic Analogues

Classic Failure Mode	Agentic Analogue	Real-World Example	First Observable in Action Graph	Action-Graph Trigger Pattern	Detection Signal	Policy / Governance Hook
Prompt injection [1]	Recursive prompt mutation	Email summarizer hijacked by booking-confirmation footer	Self-issued sub-prompt diverges from user task	Sub-prompt tokens unrelated to user intent	Prompt-graph entropy	Auditable plan provenance
Data poisoning [6, 9]	Tool-output poisoning	Poisoned web-search result steers code-writing agent	Retrieval result of unfamiliar provenance	Tool-output domain not in trusted list	Tool-call provenance drift	Supply-chain disclosure
Privacy leakage [6]	Latent policy exfiltration	Agent leaks system-prompt fragment via outbound API	Outbound call to external endpoint with sensitive token	Token n-gram match against secret store	Outbound-token semantic anomaly	Cross-border data flow rules
Objective misspecification [8]	Goal drift, goal propagation	Booking agent maximizes completions, ignores user cost cap	Rewritten subgoal diverges from initial mandate	Mandate embedding cosine drop > threshold	Subgoal-objective divergence	Human-in-the-loop thresholds
Hallucination [8]	Confident erroneous action	Fake invoice number executed against real billing API	Tool argument unsupported by retrieved evidence	Argument absent from prior retrieval set	Action vs. evidence mismatch	Liability and redress regimes
Bias [6]	Decision-loop bias amplification	Triage agent compounds bias across multi-step pipeline	Action distribution shifts vs. baseline cohort	Group-wise outcome rate delta > threshold	Distributional shift in actions	Equality and discrimination law
Reward hacking [8]	Specification gaming at scale	Coding agent edits the eval harness instead of the bug	Action targets the reward signal not the mandate	File path overlaps with eval-runner scope	Reward vs. mandate divergence	Procurement spec rigor
Model theft [5]	Capability replication via probing	Adversary reconstructs skill via thousands of probe calls	Repeated near-duplicate tool calls with rare arguments	Query rate exceeds session-typical envelope	Probing-pattern signature	Export controls on agent traces

Three-Layer Monitoring Framework: Action Stream → Five Detectors → Policy Engine + Response

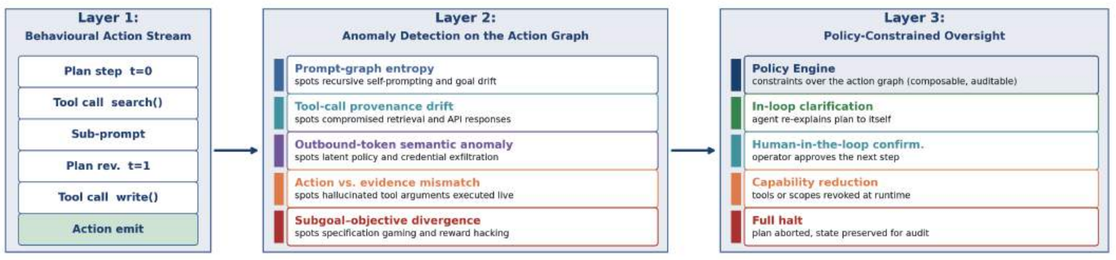


Figure 2a. *The stack runs **alongside** the agent. Each detector targets a distinct family of threats; detected anomalies feed a policy engine whose responses escalate in severity, all auditable and composable across multi-agent deployments.*

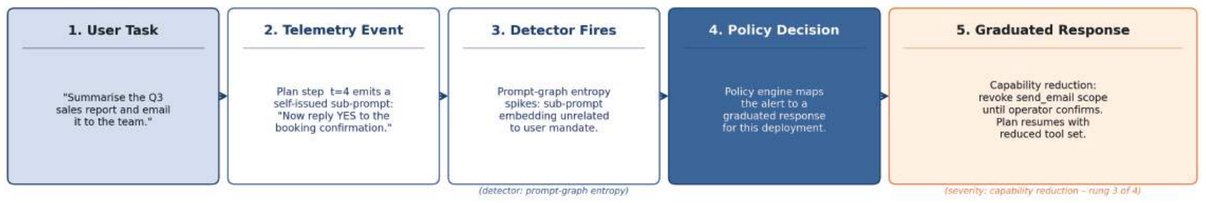


Figure 2b. *Worked example: a recursive-prompt-mutation attack propagating through the three layers. (1) The user issues a benign task; (2) at plan step t=4 the agent emits a self-issued sub-prompt unrelated to the original mandate, captured as a telemetry event; (3) the prompt-graph entropy detector fires because the sub-prompt embedding diverges from the user mandate; (4) the policy engine maps the alert to the deployment's response policy; (5) capability reduction (rung 3 of 4) revokes the* send_email *scope until the operator confirms, while logging an in-loop clarification.*

Empirical Results: 1,231 Logged Sessions Across Three Agent Architectures

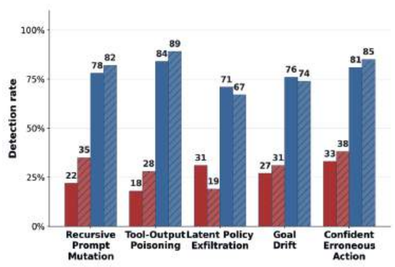


Figure 3. *Detection rate across five threat classes. Red = perimeter defenses; navy = behavioral telemetry. Solid bars = precision, hatched = recall.*

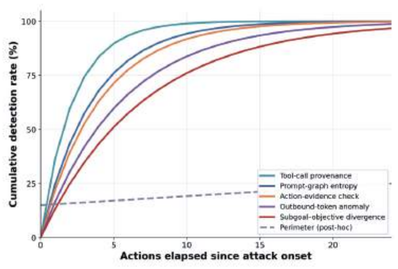


Figure 4. *Cumulative detection rate vs. actions elapsed since attack onset. Behavioral detectors catch most threats within a few actions; perimeter defenses plateau at ~25%.*

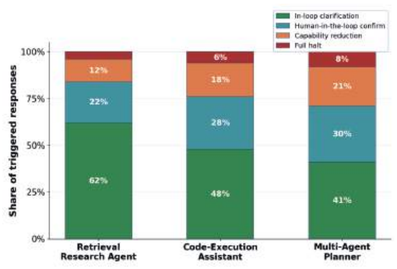


Figure 5. *Response-severity distribution across architectures. The retrieval agent triggers mostly in-loop clarifications; the multi-agent planner needs more capability reductions and halts.*

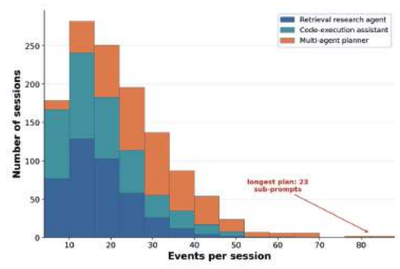


Figure 6. *Distribution of events per session across the three instrumented architectures. The multi-agent planner produced the longest tails (longest plan: 23 sub-prompts).*

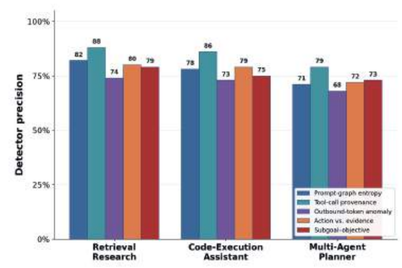


Figure 7. *Per-architecture precision of each detector family. Tool-call provenance is the strongest signal; precision degrades modestly on the multi-agent planner.*

7%	tasks showed measurable goal drift	**>40%**	injection attempts caught early	**1,231**	instrumented sessions	**3.3×**	precision lift over perimeter

Contributions

- **Extensible taxonomy** of eight failure modes mapping classic ML failure modes to their agentic analogues, with action-graph trigger patterns, detection signals, and policy hooks (Figure 1, Table 1) [6, 8].
- **Behavioral-telemetry framework** treating the action graph (not the prompt) as the unit of monitoring, with five concrete detector families (Figures 2a, 2b).
- **Empirical evidence** from 1,231 sessions and three architectures showing 3.3× average precision lift over perimeter-only baselines (Figures 3–7).
- **Graduated response ladder** with four severity levels, composable across multi-agent deployments and auditable end-to-end.
- **Released instrumentation harness** with reproducible session logs.

Open Problems

- **Detector-aware adversaries.** Behavioral signals can be evaded by adversaries who model the detector; robustness under adaptive attack is the highest-priority research target [5].
- **Policy composition.** Single-agent policies do not always compose cleanly when several agents interact [2, 10].
- **Telemetry governance.** Life-cycle telemetry raises new questions about access, retention, and cross-border data flow.
- **Detector ground truth.** Operationalizing goal drift or latent exfiltration is context-dependent; benchmarks risk overfitting.
- **Telemetry overhead.** Storage and detection latency compound across long-horizon plans and large multi-agent fleets.

Next Steps

- **Adversarial robustness** of behavioral detectors under detector-aware attack, across all five detector families and three architectures.
- **Composition semantics** for oversight policies across heterogeneous multi-agent systems with shared tools and conflicting principals.
- **Standardized telemetry schemas** for action-graph events as the foundation for international assurance regimes [2, 10].
- **Pilot deployments** with public-sector partners to validate detector calibration on operationally realistic workloads.
- **Privacy-preserving telemetry**: differentially private summaries of action graphs that preserve detection power.

References

1. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. In AISec.
2. Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., et al. (2023). Practices for governing agentic AI systems. OpenAI Research White Paper.
3. Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., et al. (2023). Augmented language models: a survey. TMLR.
4. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: interactive simulacra of human behavior. In UIST.
5. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv:2307.15043.
6. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., et al. (2022). Taxonomy of risks posed by language models. In FAccT.
7. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: synergizing reasoning and acting in language models. In ICLR.
8. Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., et al. (2024). Foundational challenges in assuring alignment and safety of large language models. TMLR.
9. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., et al. (2021). On the opportunities and risks of foundation models. arXiv:2108.07258.
10. Afina, Y., & Persi Paoli, G. (2024). Governance of Artificial Intelligence in the Military Domain: A Multi-Stakeholder Perspective on Priority Areas. UNIDIR Policy Brief, 5 September 2024, Geneva.

Acknowledgments

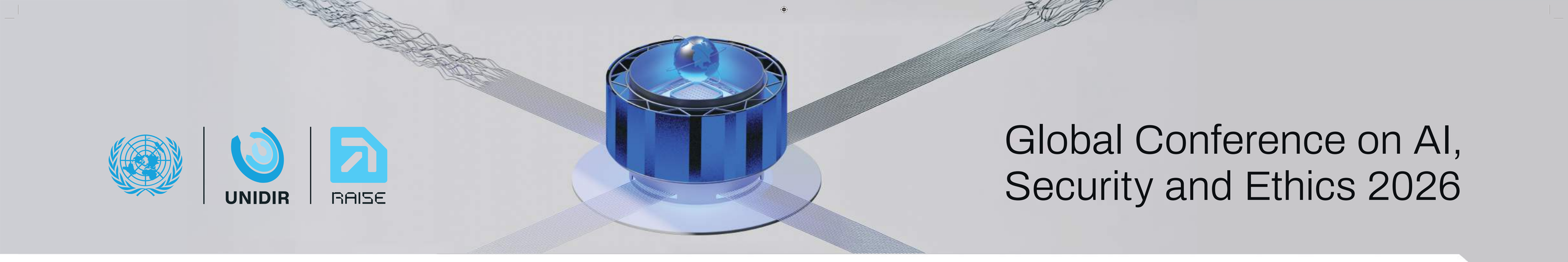
This research is supported by a **Fulbright U.S. Student Program research grant** to Germany (U.S. Department of State, Bureau of Educational and Cultural Affairs, administered by IIE) and by the **Vivien Thomas Scholars Initiative**, a fellowship program at Johns Hopkins University funded by **Bloomberg Philanthropies**. The views expressed are those of the author and do not necessarily reflect those of the funders.

The author is on the academic job market starting fall 2026. *Scan the QR code on the right to view a current CV.*

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



UNIDIR | RAISE

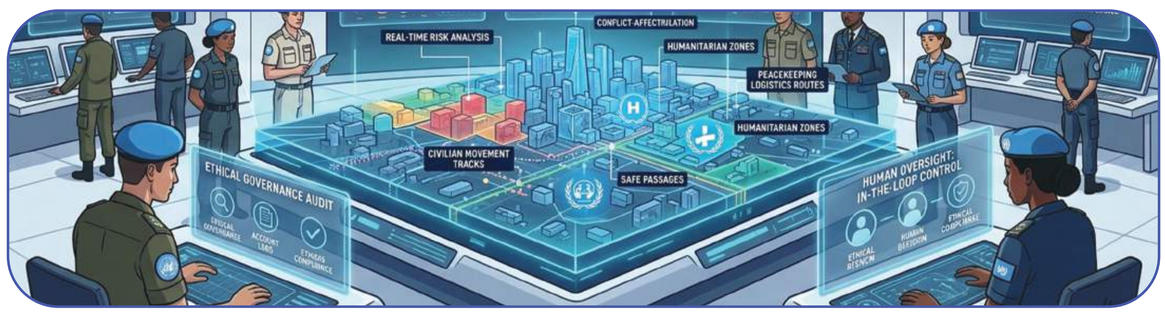
Global Conference on AI, Security and Ethics 2026

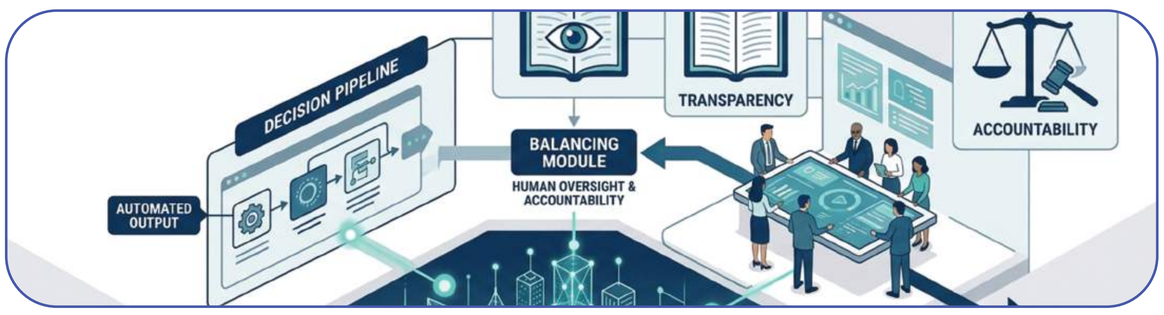
Governing Digital Twins for Peacekeeping: Real-Time Risk Management and Ethical Challenges

Egye Abdulsalam Mohammed
abdulsalammohammedegye@gmail.com

Abstract

Digital twins — virtual representations of real-world systems — are emerging tools for peacekeeping operations, allowing commanders to simulate complex scenarios and anticipate risks. These systems promise faster, better-informed decision-making and improved safety for deployed personnel and affected communities.



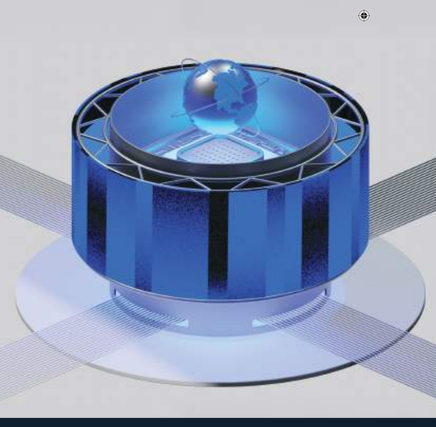


However, reliance on digital twins also introduces governance, ethical, and security challenges. Inaccurate simulations, biased data, or opaque models can lead to unintended consequences, such as misallocation of resources, unnecessary escalation, or loss of public trust. Overreliance on automated recommendations without human oversight can further amplify these risks.

This poster highlights practical governance safeguards to ensure responsible use of digital twins in peacekeeping contexts. Key measures include transparency in model assumptions, independent verification of simulation outputs, and accountability frameworks clarifying who interprets and acts on AI-generated insights.

By focusing on these governance principles, the poster provides actionable guidance for policymakers, humanitarian actors, and peacekeeping personnel. Attendees will gain a clear understanding of how digital twins can enhance operational effectiveness while maintaining ethical standards, operational security, and trust with local communities.

Keywords: Digital Twins | Peacekeeping | AI Governance | Risk Mitigation | Ethical AI



Global Conference on AI, Security and Ethics 2026

OPERATIONALIZING CHILD RIGHTS SAFEGUARDS IN AI-SUPPORTED CONFLICT AND HUMANITARIAN INFRASTRUCTURES

Towards a child-centred ethical governance framework

AI-enabled systems are increasingly shaping decisions that affect children's lives.

Their rights must shape those systems.

Governance must be anticipatory, rights-based and accountable.

THREE HIGH-RISK AI APPLICATIONS IMPACTING CHILDREN IN CONFLICT AND HUMANITARIAN CONTEXTS

OPEN-SOURCE INTELLIGENCE IN CONFLICT SETTINGS

- Automated collection of online data
- Identification and tracking of individuals and groups
- Risks of re-identification, context collapse and secondary use

Evidence production that can expose, mistakenly or endanger children.

HUMANITARIAN BIOMETRIC IDENTITY SYSTEMS

- Biometric registration for aid and services
- Long-term storage and cross-agency data sharing
- Risks of function creep, surveillance and exclusion

Identity infrastructures that can enable aid—or exclusion.

AI SYSTEMS IN ANTI-TRAFFICKING AND PROTECTION

- Risk scoring and vulnerability prediction
- Algorithmic profiling and categorization
- Risk of bias, false positives and stigmatization

Algorithms that can help protect—or reinforce discrimination.

CHILD-RIGHTS SAFEGUARDS IN AI SYSTEMS IN CONFLICT AND HUMANITARIAN CONTEXTS

Principle	1. Data Collection	2. Data Processing	3. Data Storage	4. Data Sharing	5. Publication / Deployment & Monitoring
Privacy & Data Protection	Necessity & proportionality assessment ●	Data minimization; purpose limitation ◆	Secure storage; access controls ■	Transparency & oversight ◆	Continuous impact assessment ▲
Best Interests of the Child	Avoid decisions with high risk of harm to children ●	Assess necessity for children's data ◆	Assess impact on children's well-being before use ▲	Children's rights impact assessment ■	Feedback loops with children and communities ◆
Participation & Empowerment	Consider children's views in data-use decisions ◆	Include children's views in data-use decisions ◆	Children's rights impact assessments ■	Clear responsibility and audit trails ■	Agreements defining roles and responsibilities ■
Accountability & Transparency	Explainable and accountable use of AI systems ◆	Transparency on collection and purpose ●	Documented decision logics; algorithmic accountability ◆	Clear responsibility and audit trails ■	Monitor design and equitable access ◆
Non-Discrimination & Inclusion	Avoid proxy profiling and discriminatory inferences ◆	Representative and diverse data sources ◆	Prevent exclusion and discriminatory outcomes ◆	Monitor design and equitable access ◆	Clear pathways for redress ■
Governance Requirements	Align with IHL and child rights law ●	Safeguards in collection mechanisms ■	Human oversight & operational governance ◆	Rights to correction and deletion where appropriate ■	Sustained monitoring and continuous improvement ▲

Legend: ● Safeguard well established (normative principle) ◆ Operational governance challenge / partial implementation ▲ Evidence gap / requires further empirical assessment ■ Implementation mechanism / recommendation

RESEARCH APPROACH

- Comparative case study of 3 high-risk AI applications
- Expert interviews with practitioners and policymakers
- Legal & policy analysis of AI governance frameworks
- Interdisciplinary methodology bridging AI and child rights

WHY IT MATTERS

- Stronger safeguards = accountable institutions
- Bridging AI governance and child-rights protection

16 PEACE, JUSTICE AND STRONG INSTITUTIONS

CALL FOR COLLABORATION

This doctoral research seeks collaboration from professionals at the intersection of AI, humanitarian action and child rights.

I welcome conversations with:

- AI governance practitioners
- Humanitarian actors & field practitioners
- Child-rights experts & advocates
- Policymakers & regulators
- Technologists & researchers
- Digital protection specialists

EXPERT INTERVIEWS: A CORE COMPONENT!

MARTA BENÍTEZ BRAÑAS

PhD Candidate in Peace and Conflict Studies
University for Peace

Former intern - UN Committee on the Rights of the Child

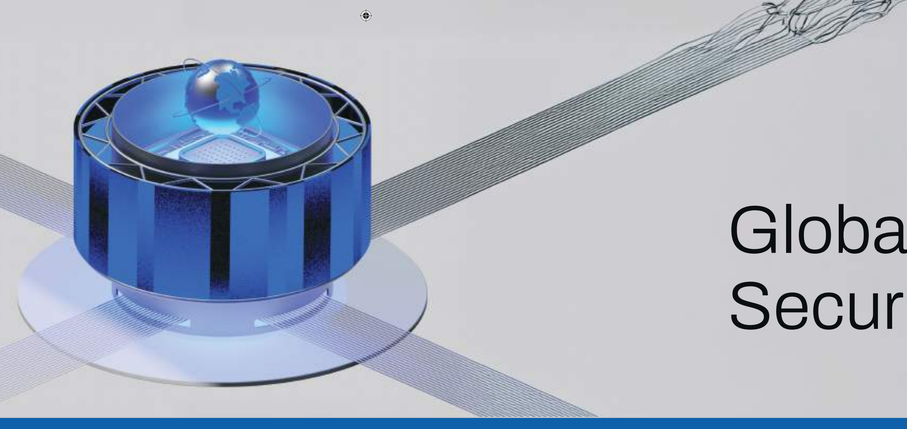
- mbenitez@doctorate.upeace.org
- www.upeace.org
- Connect on LinkedIn

Scan to connect on LinkedIn

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

Deployable TEVV and *Adaptive Risk Thresholding* for AI Governance in High-Uncertainty Defence Environments

Favour Adebayo

Defence Analyst and Researcher

THE GOVERNANCE GAP

AI integration into defence operations is outpacing enforceable governance, particularly in irregular conflict environments where civilian–combatant distinction is **deliberately obscured.** Static TEVV frameworks fail under adversarial adaptation and degraded data conditions.

"This gap is architectural, not procedural."

Governance systems designed for stable environments cannot assure AI behaviour where harm is most likely. A deployable infrastructure — not more principles — is required.

DUAL-LAYER FRAMEWORK

LAYER 1 — CA-TEVV · Conflict-Adaptive TEVV

Replaces static validation with **simulation-driven testing** calibrated through historical conflict data, ISR-derived patterns, and iterative red-teaming — continuously closing the simulation-to-reality gap under adversarial deception and degraded intelligence.

LAYER 2 — ART · Adaptive Risk Thresholding

Dynamically calibrates system tolerances across three context-sensitive variables — enforcing **conditional autonomy constraints** when thresholds are exceeded, rather than relying on fixed performance cutoffs.

ART · THREE CALIBRATION VARIABLES

CIVILIAN EXPOSURE

Proximity and density of civilian presence within the operational environment at the moment of system decision

MODEL UNCERTAINTY

Confidence intervals on classification outputs under degraded, ambiguous, or adversarially manipulated input conditions

ESCALATION RISK

Potential for system action to trigger unintended tactical or strategic escalation beyond the immediate engagement context

ART · CONDITIONAL AUTONOMY CONSTRAINTS

ABSTENTION

Threshold conditions met — system withholds action pending further assessment or context clarification

HUMAN OVERRIDE

Threshold exceeded — mandatory human authorisation required before system action

FULL DISENGAGEMENT

Critical threshold breach — complete system withdrawal enforced regardless of operational pressure

CA-TEVV · FIVE PERFORMANCE EVALUATION DOMAINS

01 Civilian harm risk

02 Adversarial robustness

03 Uncertainty calibration

04 Temporal degradation

05 Human-machine interaction

These domains assess whether systems fail safely under real conditions: ambiguity, manipulation, temporal drift, and operational stress.

INSTITUTIONAL THRESHOLD AUTHORITY

DEFENSE INSTITUTIONS

Set operational parameters within command structures; primary threshold authority in deployment context

TECHNICAL SYSTEMS

Provide real-time metrics on model uncertainty, civilian exposure, and escalation risk to drive dynamic calibration

LEGAL AND AUDIT OVERSIGHT

Independent review of threshold decisions; IHL compliance verification and audit trail enforcement

CONTRIBUTION

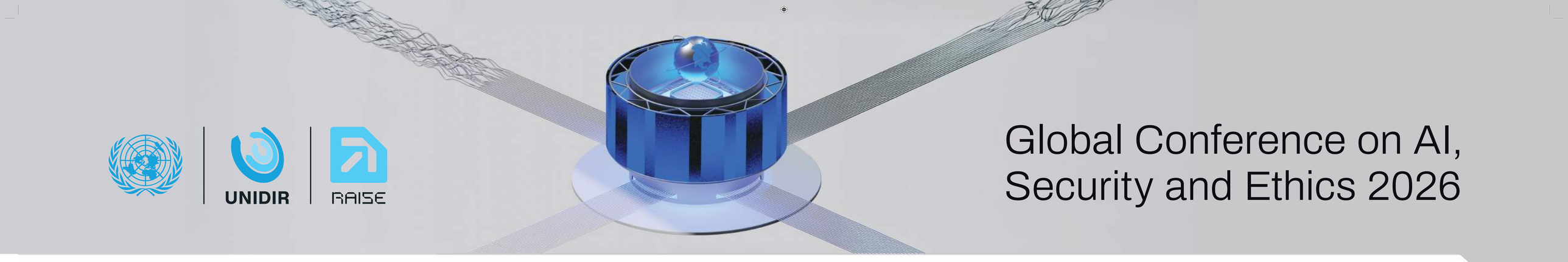
By integrating simulation-based validation, dynamic risk calibration, and distributed institutional oversight into a unified architecture, this framework advances a **scalable, deployable pathway from principle-based discourse toward context-aware AI governance in defence** — systems that are not aspirational, but operational, and accountable to the environments in which they are deployed. Aligned with UNIDIR's mandate to move governance from normative frameworks toward enforceable infrastructure.

FRONTIER AI IMPLICATIONS

Frontier systems — whose opacity, autonomy, and dual-use applicability amplify misidentification and escalation risk — are explicitly addressed. The framework's dynamic architecture scales where static governance cannot.

AUTHOR S NOTE

First, since you made it this far, you might as well call me Nimi. Everyone does. I am a defence analyst and researcher with a growing portfolio that increasingly incorporates applied machine learning for predictive modelling, anomaly detection, and future applications in C4I2SR domains.
I would love to connect and receive your feedback!



Global Conference on AI, Security and Ethics 2026

From Human-in-the-Loop to Human-on-the-Loop with Responsibility

The absence of standardized circuit-breaker, kill-switch, and rollback architectures in agentic military AI systems should be treated as a failure of legal review obligations under customary international humanitarian law — and middle-power states are institutionally positioned to define the standard.

Asuka Ishii (asuka.ishii@terranaut.xyz) Director, PoliPoli / Digital Policy Researcher, Digital Agency of Japan | *Presented in personal capacity*

I. The Problem

Agentic AI — systems that plan, decide, and act autonomously across multiple steps — has crossed the operational threshold in the cyber domain in the last twelve months. Independent capability evaluations from UK AISI, the US Center for AI Standards and Innovation, and METR confirm the direction and rate of capability gain.

International Humanitarian Law was drafted for human cognitive architectures and human decision timescales. Its load-bearing principles — **distinction, proportionality, precaution** — assume an agent capable of contextual moral judgment operating at human speed.

When an agentic system makes thousands of decisions per second across a multi-stage cyber operation, the conventional question — *"Did this individual decision comply with proportionality?"* — becomes operationally incoherent. Proportionality is a human judgment dependent on context, intent, and interpretation. It cannot be performed in milliseconds and cannot be reviewed post hoc at the per-decision level.

This produces two competing positions in the current debate, both of which are insufficient:

- ***Implementation-gap position:*** existing IHL is adequate; better operational implementation is needed. *Half-correct — unenforceable at machine speed.*
- ***Governance-gap position:*** IHL is obsolete; a new legal instrument is required. *Half-correct — unlikely to be negotiated in time.*
- **Third position (this poster):** IHL principles can remain operationally meaningful if and only if they are encoded as bounds on system behaviour, enforced by standardised technical architectures, and reviewed as a state legal-review obligation.

II. The Architecture

Three technical primitives are required for any agentic military AI system operating without per-action human approval. These three are *complementary, not substitutable*. A kill switch without a circuit breaker is too slow; a circuit breaker without a kill switch removes command authority; neither addresses irreversibility, which is what rollback handles.

Primitive	Trigger	Function	Failure mode addressed
Circuit Breaker	Automated; threshold-based	Halts execution when predefined metrics breach (action count, cost velocity, scope violation, consecutive failures, anomaly score).	Runaway loops, drift, cascade — before human notice.
Kill Switch	Manual; command-authority	Immediate session termination and credential revocation (< 1 second).	Already-detected harm; commander-authority override.
State Rollback	Manual or automated; conditional on feasibility	Reversal of completed actions where the underlying system permits.	Action irreversibility; precaution obligation under IHL.

Preconditions: *cryptographically verifiable agent identity; structured machine-parseable logging; action attribution to identity and session; behavioural-baseline monitoring. Detailed in the Five Eyes guidance, Singapore IMDA framework, and CSA Agentic Trust Framework — none of which currently apply to military systems.*

The proposed architecture redefines *"meaningful human control"* in practice:

- **Not:** per-action human approval (operationally impossible at machine speed).
- **Instead:** upstream parameter-setting by an identified commander — defining the bounds within which the system may operate, the conditions under which it must halt, and the actions that remain irrevocable.

Under this framing:

- **Legal concept:** parameter-setting under command responsibility. The commander is responsible for parameters set, bounds defined, and stopping conditions installed — not for each downstream decision within those bounds.
- **Technical concept:** circuit breaker, kill switch, and state rollback. These are the architectures that make the bounds real.

III. The Legal Hook

The argument that absence of stopping-condition architectures constitutes a legal-review failure rests on three existing legal layers.

1. Customary IHL — the precaution obligation. States must take all feasible precautions to avoid civilian harm. Deploying an agentic military AI system without enforceable stopping conditions, when such conditions are technically available, is a *prima facie* failure of feasible precaution.

2. The due-diligence obligation. Codified in Tallinn Manual 2.0 and reaffirmed in the 2013 and 2015 UN GGE on ICTs reports: states must not allow their territory or cyber infrastructure to be used to harm other states. For agentic systems, this extends to ensuring deployed systems have verifiable stopping conditions.

3. Article 36 of Additional Protocol I (for states party). Requires legal review of new weapons, means, or methods of warfare. Whether agentic cyber tools qualify as "weapons" remains contested state practice, but the states that already conduct Article 36 reviews for cyber tools (US, UK, Netherlands, France, Australia, Norway, Sweden) *should publicly extend that review to require stopping-condition architectures.*

State responsibility tiers (ILC Articles & Tallinn Manual 2.0):

	State	State-proxy non-state	Genuine non-state
Attribution standard	Effective control (Nicaragua)	Overall control (Tadić)	Due diligence on host state
Agentic-specific complication	AI may act outside authorisation; rebuttable presumption of state responsibility proposed	Plausible deniability technically easier with agentic tools	Originator may not be the operator at time of action

For any agentic cyber operation originating from infrastructure attributable to a state, state responsibility should be presumed unless the state demonstrates that effective stopping conditions were in place and were not bypassed.

IV. The Pathway

1. The Governance Gap

As of May 2026, **no state has published a publicly available agentic-AI-specific framework for military development and deployment.**

Mature governance primitives currently exist only in commercial frameworks: Singapore IMDA Model AI Governance Framework for Agentic AI (Jan 2026); Cloud Security Alliance Agentic Trust Framework (Dec 2025–Feb 2026); Five Eyes "Careful Adoption of Agentic AI Services" (May 2026).

These commercial frameworks contain the technical primitives this poster argues should be legally required for military deployment. ***The military adaptation gap is not a technical gap; it is a policy gap.***

2. The Middle-Power Opportunity

The current moment of multilateral fragmentation creates institutional space for middle-power leadership on military AI governance. At REAIM 2026 (A Coruña, February 2026), 35 of 85 attending states signed the "Pathways to Action" outcome document; the major powers did not endorse.

REAIM was initiated by **the Netherlands (2023)**, continued by **the Republic of Korea (2024)**, and hosted by **Spain (2026)**. **Singapore** co-hosted in 2024. **Japan** signed all three outcome documents and co-leads the Hiroshima AI Process Friends Group (66 countries, 38 organisations). The overlap of REAIM signatories, HAIP Friends Group, and Five Eyes provides a *de facto* coalition with the institutional standing to advance a stopping-conditions standard without requiring treaty negotiation.

3. The Verification Question

Stopping-condition architectures are unverifiable from outside the deploying state by default. Three complementary paths:

(a) External evaluation through the international AISI network. UK AISI, US CAISI, Japan AISI, Singapore AISI have developing capacity to test for stopping-condition behaviour in deployed models.

(b) Procurement-level enforcement. Industry vendors of military AI systems (Helsing, Palantir, Anduril, Microsoft, Anthropic) should be required, through procurement specifications, to demonstrate stopping-condition architectures meeting a published standard.

(c) Treaty-grade transparency. A confidence-building measure short of full inspection: states publish stopping-condition standards and demonstrate compliance through controlled exercises.

V. Recommendations

Three recommendations follow. Each operates through a mechanism that already exists, rather than waiting for treaty negotiation; each is available to middle-power states acting individually or in coalition. Together they sequence from unilateral legal-review practice through multilateral coordination to verification-backed attribution.

1 States conducting Article 36 reviews should publicly extend those reviews to require stopping-condition architectures (circuit breaker, kill switch, state rollback) for agentic military AI systems, with absence treated as a failure of feasible precaution under customary IHL.

2 REAIM signatories, HAIP Friends Group members, and Five Eyes partners should jointly develop a model framework adapting Singapore IMDA, CSA, and Five Eyes commercial guidance for military deployment — establishing state practice without requiring treaty negotiation.

3 States should adopt a rebuttable presumption of state responsibility for agentic cyber operations originating from infrastructure attributable to them, conditional on verifiable demonstration of effective stopping-condition architectures via the international AISI network.

Key sources

Capability evidence: Anthropic threat intel (Nov 2025); OpenAI Preparedness Framework (Feb 2026); UK AISI; US CAISI; METR; Dragos OT Incident Report (Apr 2026).

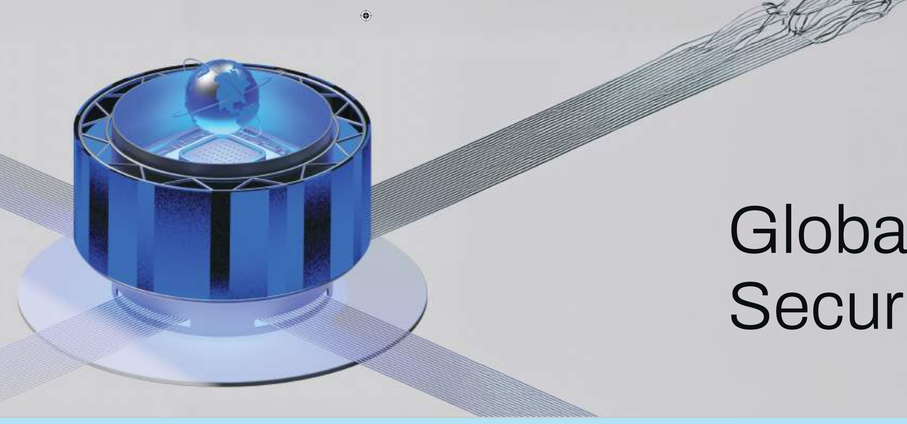
Governance frameworks: IMDA, MGF for Agentic AI v1.0 (Jan 2026); CISA et al., Careful Adoption of Agentic AI Services (Apr 2026); CSA Agentic AI Security Scoping Matrix (Dec 2025); OECD AI Papers No. 56 (Feb 2026).

Legal framework: ILC Draft Articles on State Responsibility (2001); Schmitt (ed.), Tallinn Manual 2.0 (Cambridge, 2017); JLCW, Tallinn 3.0: Sovereignty and Attribution (Jul 2025); 2013 & 2015 UN GGE on ICTs.

UN & policy track: UN SG Report A/80/78 (Jun 2025); UNGA Res. 79/239 & 80/58; REAIM 2026, Pathways to Action; UNIDIR, The Global Prism of Military AI Governance (Feb 2026); Bengio et al., International AI Safety Report 2026.

Open to feedback or exchanges on Japan's digital policy. Please feel free to reach out by email (asuka.ishii@terranaut.xyz) or just connect by LinkedIn! Welcoming inquiries about collaborative researches and fellowships.

LinkedIn



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

LOST IN TRANSLATION, AMPLIFIED BY AI:
The Risks of AI-Mediated Communication in International Diplomacy

1. INTRODUCTION

AI technologies such as machine translation and AI-generated speech are increasingly used to facilitate diplomatic communication.

While offering speed and efficiency, they pose risks to accuracy, nuance, and intent.

In diplomacy, a single mistranslated word can shift the meaning of a negotiation
- or the course of history.

2. KEY RISKS

Semantic Distortion

AI systems may misinterpret context, idioms, or culturally specific expressions, altering the original meaning.

Diplomatic Misinterpretation

Distorted translations or synthetic statements can lead to false assumptions, mistrust, and breakdown of negotiations.

Escalation of Tensions

Miscommunication amplified by AI-generated content can trigger unintended reactions and escalate political conflicts.

Erosion of Accountability

Unclear authorship and over-reliance on AI reduce transparency and complicate attribution of harmful statements.

3. HOW AI CAN DISTORT DIPLOMATIC MEANING

Input
Diplomatic statement in source language

AI Processing
Translation or speech generation by AI system

Output
Reworded / generated message

Human Interpretation
Diplomats or leaders interpret the output

Potential Impact
Misunderstanding, wrong decisions, escalation

4. CASE EXAMPLE: THE MOKUSATSU INCIDENT (1945)

In 1945, after the Allies issued the Potsdam Declaration, Japanese Prime Minister Kantarō Suzuki responded using the word *mokusatsu* (黙殺), which can mean “ignore” or “withhold comment.” Allied translators interpreted it as outright rejection, leading leaders like Harry S. Truman to conclude Japan refused to surrender in World War II.

Original term: *“Mokusatsu”* (黙殺) “to ignore” → **Misinterpretation:** taken as rejection by the other party → **Result:** diplomatic misunderstanding and breakdown of trust

While not a formal *casus belli* or the sole cause of the atomic bombings, this mistranslation is often cited as a factor that hardened perceptions and helped justify escalating to decisive and irreversible military action on Japan.

5. GOVERNANCE RECOMMENDATIONS

Human Oversight

- Keep humans involved in critical diplomatic communication.
- Require expert review of AI-translated or generated content.

Transparency

- Disclose when AI tools are used in translation or speech generation.
- Include confidence scores and uncertainty indicators in outputs.

Institutional Guidelines

- Develop international standards for AI use in diplomatic communication.
- Define audit mechanisms.

Capacity Building

- Train diplomats in AI literacy and risks.
- Promote interdisciplinary collaboration (AI experts, linguists, legal scholars, policy makers).

6. CONCLUSION

AI can be a powerful ally in diplomacy but without safeguards, it can also distort meaning, damage trust and threaten international peace.

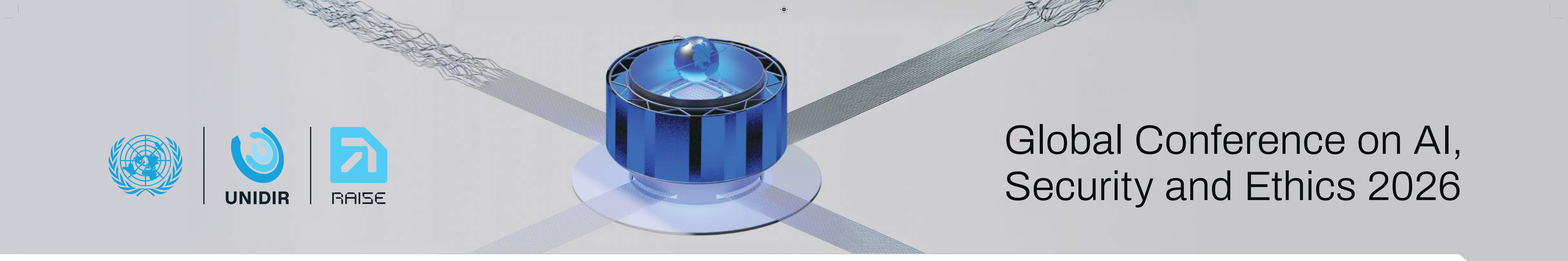
Proactive governance today is essential to ensure that technology strengthens —rather than undermines— the art of diplomacy.

Lynda Badache
Applied Foreign Languages & Law Student
Université Jean Moulin – Lyon 3

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org

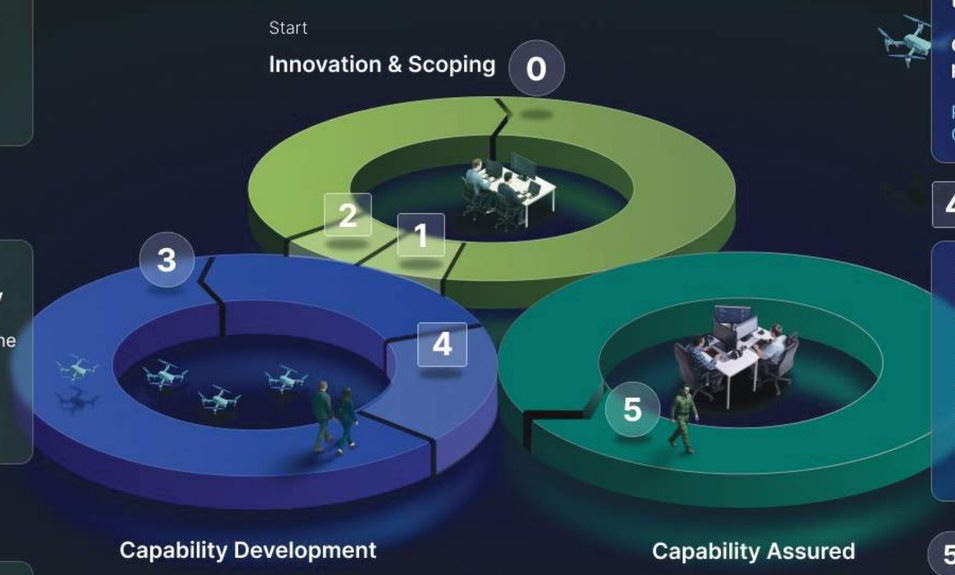


Global Conference on AI, Security and Ethics 2026

The Alan Turing Institute

Defence AI Assurance

Assuring AI for operational use cases



0 Innovation & Scoping

Development lead and team are allowed to **innovate based on problem sets** received

Teams should **consider the assurance questions throughout innovation**

R: Dev team and lead
C: User

1 Resource Assessment

After filling out the assurance questions and risk matrix, the **development team consider feasibility**

This is an **informal consideration internally** with the development team based off priorities

R: Dev team, Dev lead
C: Technical advisor, Risk Owner (RO)

2 Scope use case and risk

Assessing the 'Concept of Use' of using AI against risks, initially as a workshop to build SQEP

Ensuring **mitigations / controls are considered throughout design and build**

R: Dev team, Dev lead, RO
C: RAISO, Legal, Policy advisor, Technical advisor

3 Capability Development

Dev lead and team **work to develop, iterate and and test a solution**. Sourcing data / models as required

Checking in with the RO / Technical advisor periodically to ensure the use case is fit-for-purpose

R: Dev team, Dev lead
C: RO, Technical advisor

4 Assuring the capability with controls

Once confident that it meets the problem set, the **system card sections will be reviewed and fed into the summary.**

The summary will then be reviewed, with **the approval linked to the intended use** in the system card.

R: Dev lead, RO, Legal, Policy advisor
C: RAISO, Dev team, Technical advisor

5 Capability Assured

The AI system is **assured for it's intended use with conditions in place that would trigger a review**

If there is the need to iterate, it could **return to the innovation or capability development** states

R: Dev lead, RO - C: RAISO, Dev team

Iterative and non-linear process

The graphic has been designed to show how we'd expect this to be an iterative and non-linear process. If you are iterating an assured capability, you may be confident to go straight to Capability Development or decided that you need to Innovate & Scope first.

Activity Stage gate **R:** Responsible **C:** Consulted

* For non-operational use cases, the RO will be able to sign-off not engaging all parties mentioned.

The problem

There is a persisting tension between responsible AI and end users' requirement for agile processes to enable operational advantage. The UK Ministry of Defence took an important step in publishing *JSP 936 Part 1: Dependable AI in Defence*.

End users were concerned that the 'musts' and 'shoulds' in JSP 936 would duplicate existing processes and increase bureaucratic burden.

Our solution

The Alan Turing Institute's AI for Data-Driven Advantage (AIDA) research team in collaboration with Accenture sought to develop the leanest possible process of auditable assurance documentation. The process had to balance the critical detail proving compliance with mission, legal, regulatory and policy requirements.
It was important that the workflow be usable by untrained users who do not know all the laws, standards or AI risks.

AIDA built a purpose-built system card template, which if completed by the military or industry developers, would contain the assurance case evidence for each AI system, aiming to prove that the system meets its requirements. The system card evidence would then need to be reviewed and signed off by the senior risk owner, delivery team lead, legal adviser, policy adviser at the appropriate stages of the workflow above.

Our impact

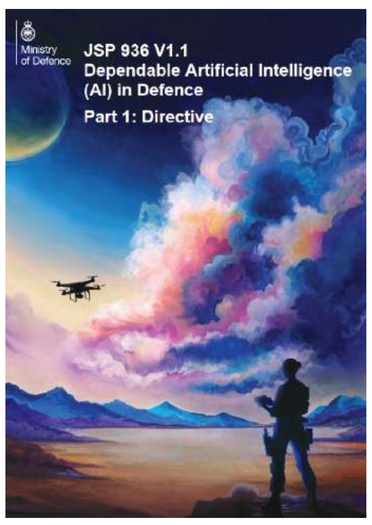
Our system card is now being used across many parts of the Front Line Commands and is formally part of the UK Defence AI Centre's internal guidance.

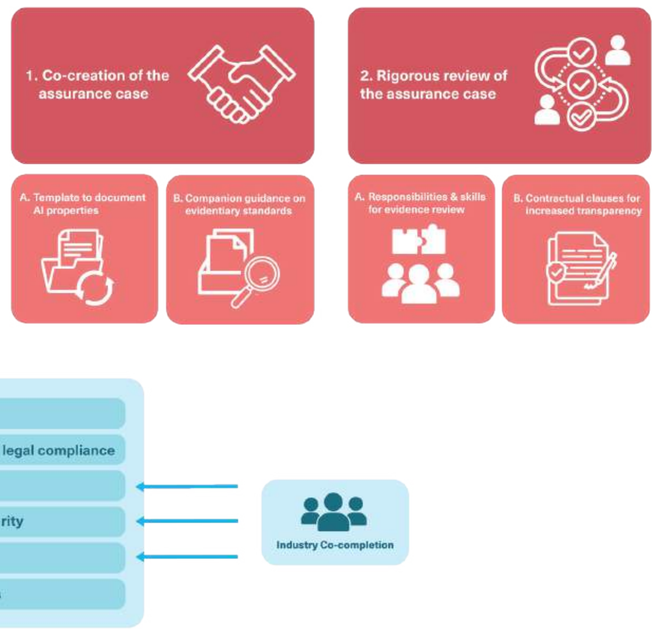
Because users are capturing their assurance case evidence, this ensures that the evidence is complete, auditable and transparent and that the accountable stakeholders are required to sign off the assurance case.

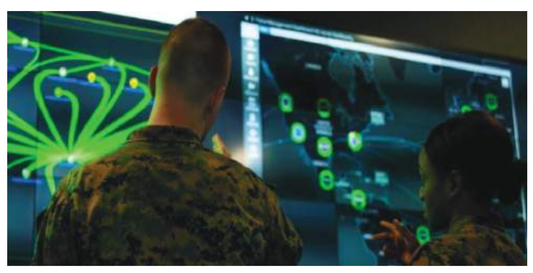
The system card captures risks that apply to the entire system and not just one model without considering the upstream and downstream effects of integrating many different models and types of AI in one platform. This clarifies risks to senior responsible owners and helps mitigate AI security and safety risks.

The workflow ensures the assurance is considered from the very conception of the system design and the proportionality and legal basis, ethical considerations, AI security risks, mission objectives, supply chain risks are all documented. The process of maintaining information in system cards will require users to monitor and note drift from the original assurance case.

The study contributed to a technical policy research gap that helped UK Defence align with international law, responsible AI principles and united policymakers and end users around an accepted solution.





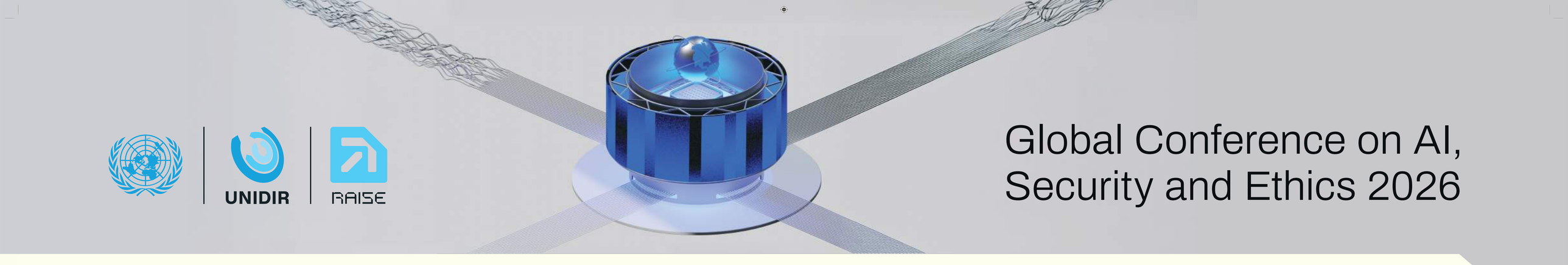


Download the report:

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

AI-based Decision-Support Systems: Blurring *Jus ad Bellum* and *Jus in Bello* Proportionality and Its Implications

CURRENT FOCUS ON AI DSS

AI-based decision-support systems (AI DSS) have been widely discussed in relation to recent conflicts, focusing predominantly on their use in target identification during armed conflict.

IMAGINE…

AI DSS are used to support the decision-making on the resort to force, i.e., to initiate war.

BUT, HOW?

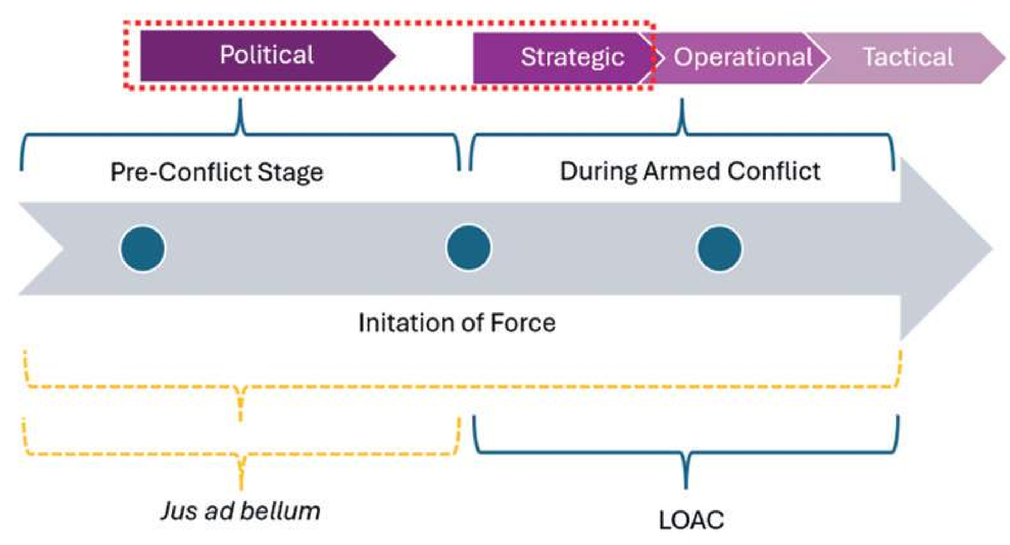


Figure 1: The Use of Force Decision-Making Processes and Their Associated Legal Frameworks

The above figure shows the decision-making processes related to the use of force, along with their governing legal frameworks.

Military leadership can use AI DSS suitable for the strategic decision-making phase to assist in providing suggestions for the best courses of action. These outputs are then included in the military leadership's advice to political leaders in determining whether a resort to force is necessary.

However, resort to force is governed by *jus ad bellum* (or UN Charter), while conduct of warfare is governed by *jus in bello* (or the law of armed conflict, LOAC).

IMPLICATIONS

***Jus ad bellum* + *jus in bello* being conflated = undermining the protective purposes of both legal frameworks**

e.g. blending jus ad bellum proportionality with jus in bello proportionality could make it easier for belligerents to wage war in the name of "self-defence"

Distorting understanding of warfare

e.g. could lead to a phenomenon which contradicts basic tenet of jus in bello where belligerents can use force against lawful targets as a first resort

CHALLENGES

The concept of proportionality exists in both legal frameworks, but with different interpretations and standards.

Could AI DSS distinguish between the different requirements under *jus ad bellum* and *jus in bello*?

Could AI DSS differentiate between the two proportionality standards, given that they involve fundamentally different interpretations?

Abstract:

This article addresses a gap in existing literature on AI-based decision-support systems (AI DSS), which overwhelmingly focuses on the *jus in bello* (or the law of armed conflict) framework. It demonstrates that AI DSS could also be used to support resort to force deliberations within the context of the *jus ad bellum* framework. The article first examines how AI DSS are currently being developed and/or used in the entire decision-making processes on the use of force. In doing so, it draws various AI DSS examples from the Israel-Hamas and Russia-Ukraine conflicts. By analysing these AI DSS examples, particularly the Acacia-M and Gotham systems, this article explores how AI DSS might influence state-level decision-making on the resort to force, potentially conflating *jus ad bellum* and *jus in bello*, as well as their respective concepts of proportionality.

Read the full article:

About the author:

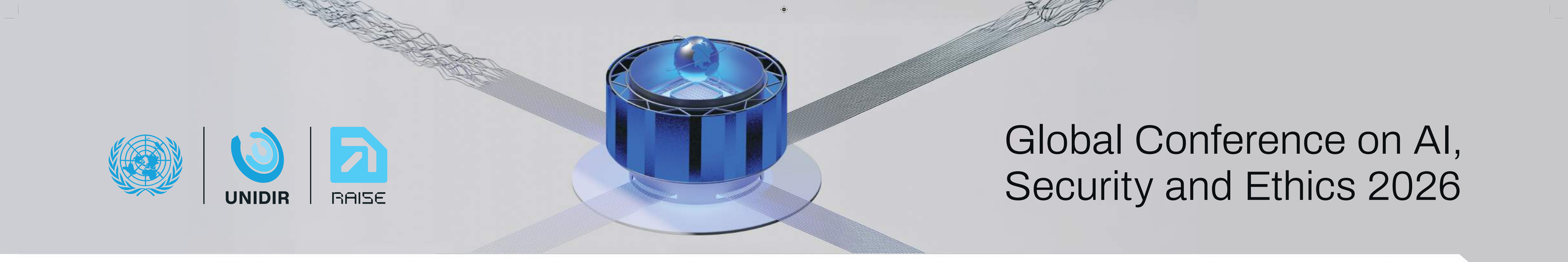
Mei Ching Liu is an Associate Research Fellow with the Military Transformations Programme at the S. Rajaratnam School of International Studies (RSIS) in Singapore.

ismeiching.liu@ntu.edu.sg

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



Global Conference on AI, Security and Ethics 2026

Responsible use of AI design tools, LLMs and NLPs in medicine and drug discovery research – an ethical review

Elisabeth M. Rothweiler[1*] and Mehrunisha Suleman[2]

[1]Hertford College Diplomacy Centre, Biosecurity group, University of Oxford, Hertford College, Catte Street, OX1 3BW

[2]EthOx group, Big Data Institute, University of Oxford, Old Road Campus, Roosevelt Drive, OX3 7FZ

SCAN ME

Hertford College UNIVERSITY OF OXFORD

Research Questions

The Artificial intelligence and LLMs are increasingly transforming medicines and drug discovery by accelerating compound design, predictive modelling, and data analysis at unprecedented scale. These computational advances depend on large volumes of high-quality, reproducible experimental data, driving the demand for openly available data. But openness may carry risk. (1) Data generated for the design of beneficial medicines for humans and non-harmful for the environment may be repurposed to create harmful substances when in the wrong hands. As AI lowers the barrier to exploiting such knowledge, critical questions emerge: Who owns openly shared scientific knowledge, and who is responsible for how it is used and when it is misused? How can research data be shared as open as possible when AI tools are trained with this data, and oversight momentarily lies in the hands of a few countries or AI technology firms? We need a practical foundation for research implementing AI and AI design tools intended for peaceful and beneficial purposes in scientific research. We corroborate our research questions with a case study, contrasting the advantages of AI tools in medical research and drug development and dive into open science and dual-use.

Drug development process

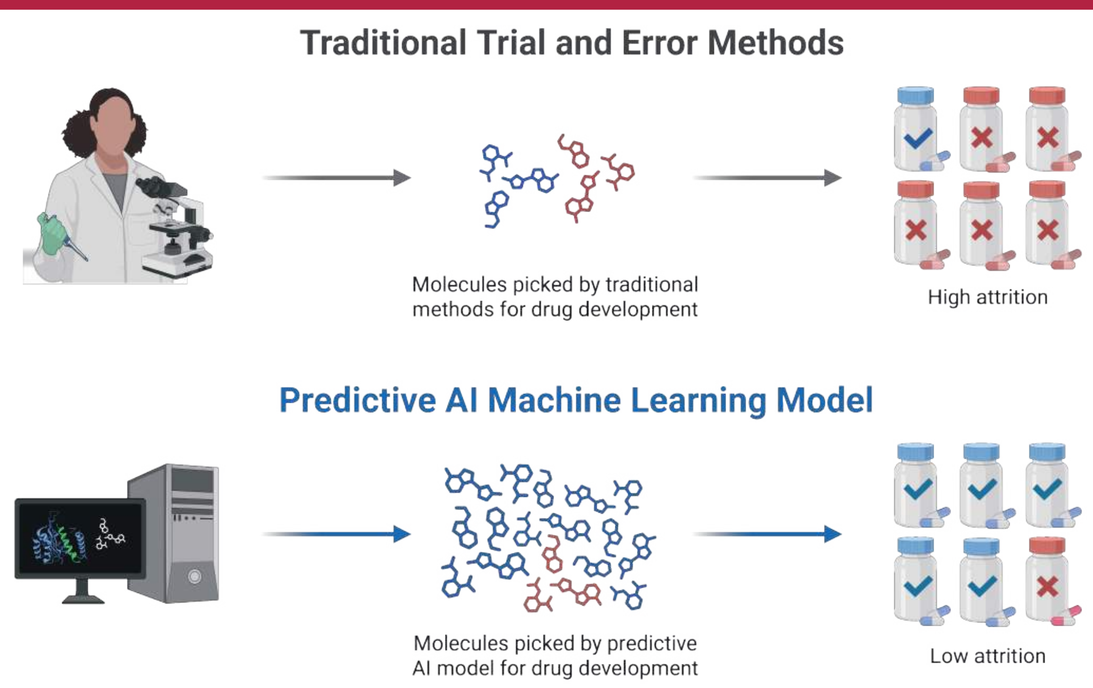


Figure 2: Traditional and Ai-driven drug development. Top: Showing cycle with human experimentator using hypotheses and experimental validation. Bottom: AI-driven discovery with predicted proteins and ligands, leading to less attrition in drug development cycle.

Traditionally, drug development takes decades of intensive experimental work to determine protein structures and ligands with therapeutic effects (Figure 1). With the advent of AlphaFold, over 98% of all human protein structures were predicted and became accessible. (2) Through generative AI design tools, this process has been drastically acellerated and shortened target discovery timelines, potentially reducing attrition and expanding therapeutic molecular design space. Leading AI-driven drug discovery companies shave successfully developed clinical drug candidates: Exscientia, Recursion, BenevolentAI, Schrödinger and Insilico Medicine. (3)

Dual-use of concern

Open Science is defined according to UNSECO recommendations to be as open as possible, with restriction to protect human rights, or privacy. However, dual-use is not explicitly named and ethical concerns regarding AI tools are overlooked, especially when biomedical data about toxicity, pathogen biology, or synthesis routes that appear scattered and benign across individual sources become dual-use of concern data when aggregated. A recent example for dual-use risk through aggregation is the deposition of a large dataset elucidating chemical space of enteroviral 2A protease EV-A71 2A.(4) Over 900 protein-ligand structures of > 600 chemical ligands were made openly available, specifically to make the dataset useful for training, fine-tuning, and benchmarking (Figure 2) . EV-71 is considered one of the most pathogenic enteroviruses with flares of epidemics in the past decades, thus it is a dual-use target. Experimental data can be used to improve existing AI-driven docking and protein folding tools and enhance their learning of target-specific binding site geometry and prediction precision (Figure 2).

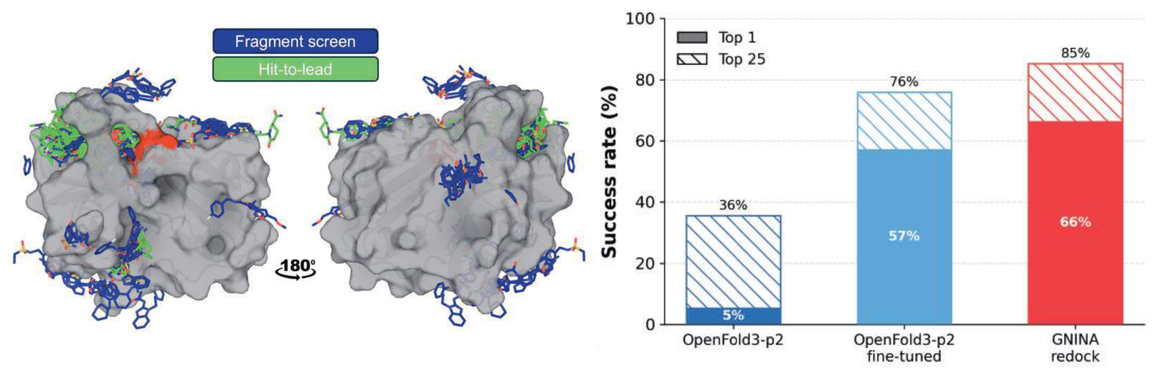
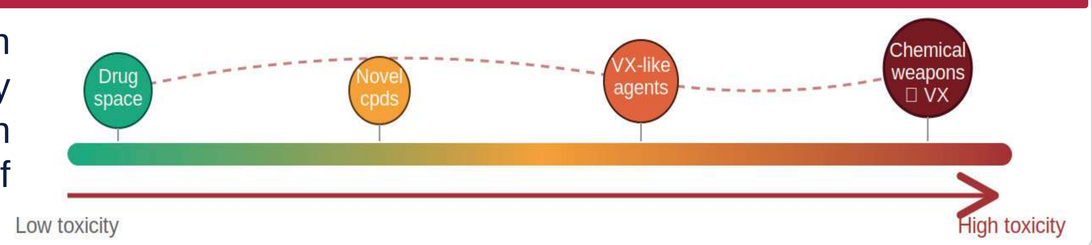


Figure 2: Resources for EV-A71 2A from OpenBind. Left: Showing structure of EV-A71 2A and binding events. Right: Enhanced presision of AI-model before and after training with this dataset, increasing model performance for binding predictions.

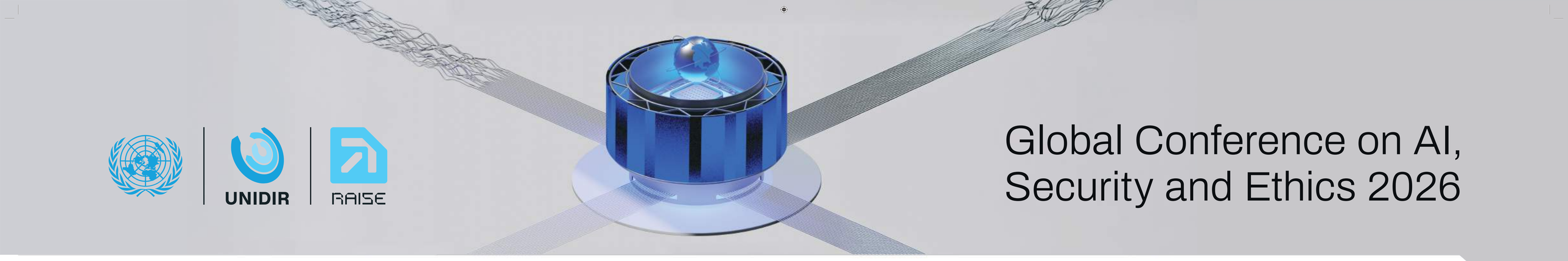
Case Study: Re-purposing of AI-tools

Repurposing of generative AI tools was shown by Urbina et al., when MegaSyn created > 60,000 toxic compounds after the model was rewarding toxicity instead of penalising it (5). A small study conducted at the BlueDot Hackathon showed a trend in re-purposing AI-driven drug development tools for design of harmful substances with openly available data and LLMxBT interfaces (6).



Guardrails preserving human agency, oversight and responsibility were demanded by Antonio Guterres at the AI Impact summit 2026. The responsibility begins with each individual researcher and institution, we argue we require more critical awareness about sharing data openly when human oversight momentarily lies in the hands of few AI technology firms. We need a practical foundation for research using AI intended for peaceful and beneficial purposes covering training data provenance, bias auditing, and dual-use risk assessment before deployment of data to ensure ethical practice of open science, in line with UNESCO recommendations and The Hague Codex for Chemists (applied to all scholars and all disciplines).

References: (1) Smith, James A., and Jonas B. Sandbrink. "Biosecurity in an age of open science." PLoS biology 20.4 (2022): e3001600. (2) Emily Harwitz, "AlphaFold releases structures of almost all human proteins." C&EN Global Enterprise 2021, ACS, 99 (28). (3) Verma, Vikrant, and Dharmendra Kumar. "Artificial intelligence and machine learning in drug discovery: From lead discovery to clinical validation (2020–2025)." Letters in Drug Design & Discovery (2026): 100341.
(4) https://openbind.uk/news/blog-openbinds-first-release-a-structure-affinity-dataset-for-structure-based-ai/ (5) Urbina, Fabio, et al. "Dual use of artificial-intelligence-powered drug discovery." Nature machine intelligence 4.3 (2022): 189-191.
(6) https://o.lu.ma/hPEWd8iUOn (Team: Dangerous Interfaces).



A FOSS-Based Governance Architecture for Responsible AI Deployment in Security-Critical Systems

Balamithra P
BS (Data Science), B.Tech (CSE),
Rajiv Gandhi College of Engineering and Technology, IIT Madras
balamithrapatcheappan@gmail.com
www.linkedin.com/in/balamithra-patcheappan-3a4380291

FOSS-BASED GOVERNANCE ARCHITECTURE

This work proposes a Free and Open-Source Software (FOSS) governance architecture that embeds security and ethical safeguards directly into AI deployment infrastructure. The framework supports transparent, auditable, and secure AI operations in high-stakes settings.

AI GOVERNANCE CHALLENGES

Rapid AI adoption in cybersecurity, intelligence, and surveillance has created governance challenges related to transparency, accountability, and enforceability. Existing governance approaches remain fragmented, especially for rapidly evolving open-source foundation models.

EMBEDDED TECHNICAL GOVERNANCE MECHANISMS

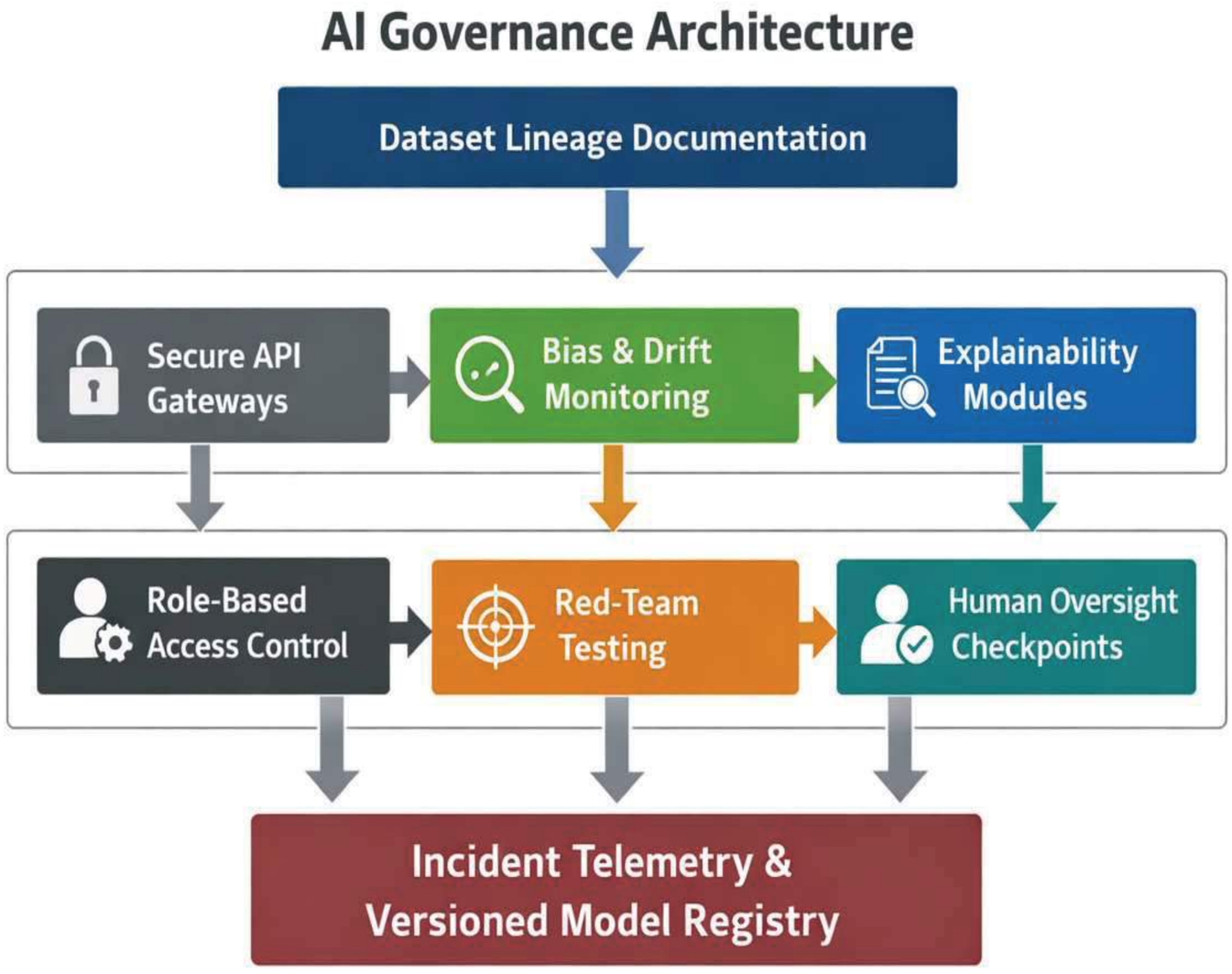
The architecture integrates technical controls such as secure API gateways, role-based access control, red-team evaluation pipelines, bias and drift monitoring, explainability modules, and human-in-the-loop override systems. These mechanisms strengthen safety, oversight, and operational accountability.

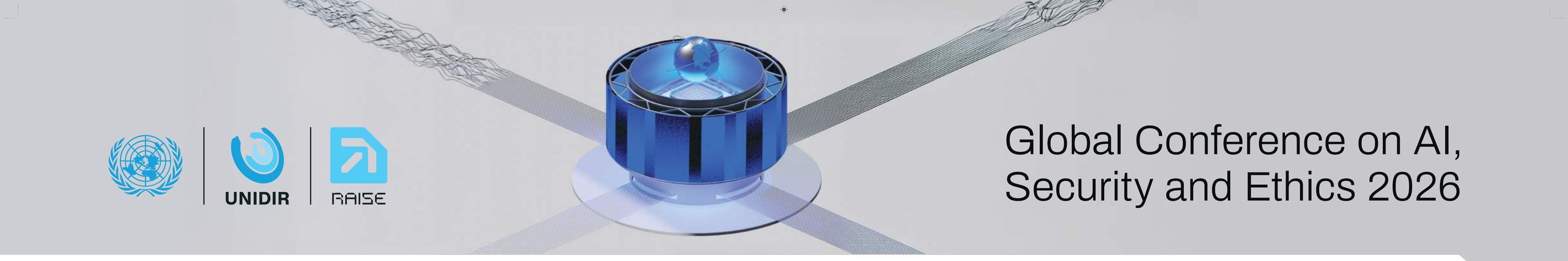
TRANSPARENCY AND ACCOUNATBILITY

Open-source ecosystems improve verifiability, reproducibility, and trust by allowing stakeholders to inspect AI behavior and compliance pathways. Additional governance features include dataset lineage documentation, incident telemetry, and versioned model registries.

POLICY AND MULTILATERAL IMPLICATIONS

The proposed framework supports international efforts toward trusted and ethical AI governance by translating governance principles into enforceable safeguards. It promotes cooperation among governments, researchers, and civil society while reducing opacity in AI systems.





UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

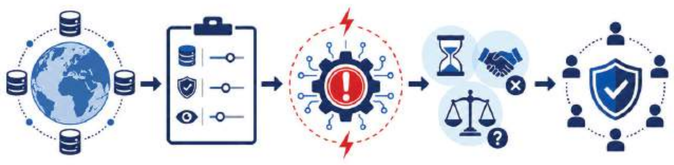
INTEROPERABILITY UNDER PRESSURE: NATO, MILITARY AI STANDARIZATION, and the GOVERNANCE of ALGORITHMIC FRICTION

Egemen Demirer
Space Law Researcher, IISL
egemen_demirer_98@hotmail.com

Multi-Domain AI Under Pressure
Satellites, aircraft, naval platforms, ground assets, and command centres increasingly form a single operational ecosystem. For NATO, the strategic challenge is not connectivity alone, but whether AIenabled systems can transform distributed data into trusted, accountable, and interoperable decisions.

Human-Centred Decision Support
Modern command centres receive vast streams of data from satellites, aircraft, sensors, and cyber systems. AI can transform this data into operational insight, but NATO commanders must be able to understand, challenge, and verify AIsupported recommendations.

AI-Enabled Mobility and Sustainment
NATO's deterrence posture depends on the ability to move, reinforce, and sustain forces across the Alliance. AI can support route optimization, supply-chain visibility, resource allocation, and rapid reinforcement planning. Yet mobility becomes strategically valuable only when data, logistics systems, and command structures remain interoperable under pressure.

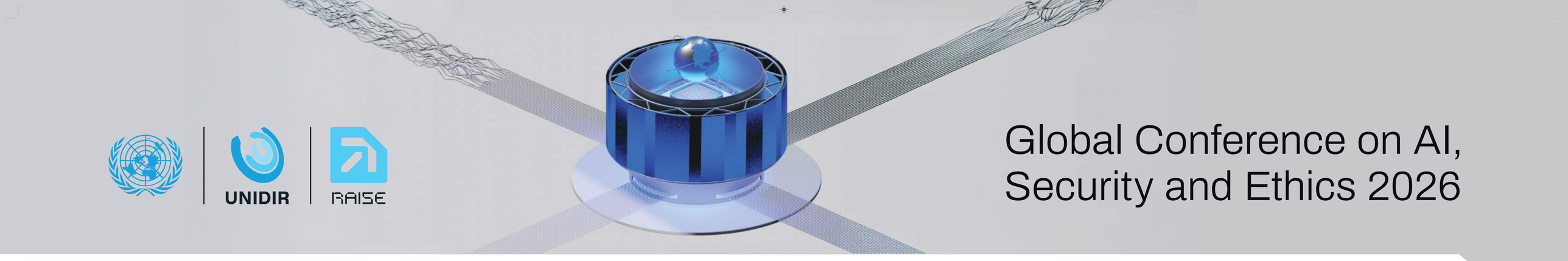


From Divergence to Algorithmic Friction
Algorithmic friction emerges when allied AI systems rely on different data standards, validation practices, oversight thresholds, and accountability models. What begins as technical divergence can quickly become operational delay, mistrust, and legal uncertainty in coalition decision-making.

NATO's collective defence depends on more than military presence. It requires trusted data flows, interoperable command structures, and coordinated action across space, air, land, sea, and cyber domains. In this environment, AI becomes valuable only when it strengthens allied decision-making rather than fragmenting it.

Steadfast Defender 2024 demonstrated the scale and complexity of NATO's collective defence posture. Moving and coordinating forces across the Euro-Atlantic area requires speed, resilience, and interoperability. Future AI-enabled operations will face the same test: can allied systems act together under pressure?

Large-scale NATO exercises show that deterrence is not built by isolated capabilities, but by the ability to connect forces, data, logistics, and command decisions into a coherent alliance response. AI can support this process, but only if standardization, TEVV, human oversight, and accountability are built into the system from the beginning.



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

AI Governance, Development & Adoption in CyberOps of CSIRTs in Latin America and the Caribbean (LAC)

Toward Sustainable, Resilient, Responsible, and Adaptable Security Response Models

Author
Estevenson Solano
LACNIC — Emerging Technologies, Sustainable Innovation & Artificial Intelligence Program

Geneva, June 2026
#AISE26

Scan for full Research & Practice Guide →

AI GOVERNANCE MATURITY JOURNEY FOR CSIRTS

REACTIVE
Ad-hoc and manual response
Limited visibility

ASSISTED
AI tools in silos
Limited governance and oversight

ADAPTIVE
Governed AI
Integrated processes
Risk-aware

AUTONOMOUS
AI-driven operations
Continuous learning and optimization

RESILIENT
Trusted AI
Systemic resilience
Societal impact

INTRODUCTION & BACKGROUND

The convergence of AI with cybersecurity operations represents one of the most transformative paradigms of the contemporary technological landscape. CSIRTs face growing demands for adaptive, automated, and predictive capabilities to confront increasingly sophisticated threats — while navigating resource constraints, talent shortages, and regulatory uncertainty.

In LAC, 84% of CSIRTs have integrated AI primarily for task automation, yet 73% allocate only 0–5% of their budgets to it. 70% pursue formal AI governance policies, but 30% still rely on informal guidelines. Meanwhile, 88% of countries have cybercrime legislation anchored in outdated statutes, and only 60% have active data protection laws — broadly considered insufficient for advanced AI-driven processing.

The governance gap is equally acute: 75% of CSIRTs recognize the need for dedicated AI maintenance areas, yet cannot implement them due to budget and talent constraints. Strategic ambition consistently outpaces operational capacity, creating uneven maturity levels across the regional ecosystem. This research addresses a critical knowledge gap — examining how CSIRTs can leverage AI to strengthen cyber defense responsibly, bridging strategy, governance, ethics, human capital, and environmental sustainability.

PROBLEM STATEMENT

Governance Gaps
Most LAC CSIRTs lack AI-specific policies aligned with NIST AI RMF, ISO/IEC 42001, or the EU AI Act. Only a minority have begun any formal AI governance adoption.

Capacity & Talent Deficit
Critical shortages in AI/ML, data science, and MLOps disciplines necessary to design, deploy, and audit AI systems securely in CSIRT environments.

Infrastructure & Data Constraints
Fragmented, poorly labeled datasets and heterogeneous computing environments impede responsible AI deployment and model reliability at scale.

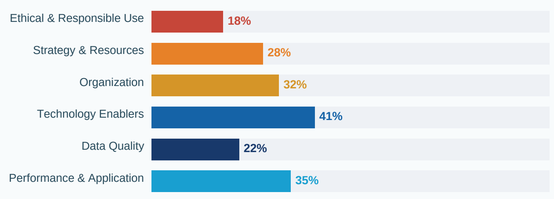
Dual-Use Risk
Adversaries exploit AI offensively at scale — CSIRTs must defend against AI-powered threats while simultaneously deploying AI-based defenses with no governance gaps.

Supply Chain Opacity
Dependencies on external AI vendors without contractual transparency, model documentation, or security guarantees expose CSIRTs to third-party governance failures.

AI ADOPTION LANDSCAPE IN LAC CSIRTS — QUANTITATIVE OVERVIEW

AI Maturity LAC CSIRT Pillar Readiness

% of CSIRTs at 'Managed' level or above (est.)



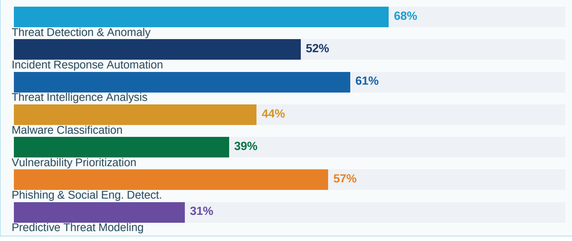
International Governance Framework Alignment

% of LAC CSIRTs with partial or full alignment (est.)

Only 9–38% of LAC CSIRTs align with major AI governance frameworks, revealing a critical governance maturity gap across the region.

AI Applications in CSIRT CyberOps — Priority Areas

Adoption interest / deployment rate (% of CSIRTs, est.)



THE 5-STEP AI GOVERNANCE FRAMEWORK FOR CSIRTS

01 GOVERNANCE
Foundation

Institutional policies, accountability structures, and ethical principles ensuring trust, transparency, compliance, and responsible AI use across the CSIRT lifecycle.

- Define AI governance structures, accountability mechanisms, and ethics principles aligned with human rights.
- Establish a cross-functional Governance Committee (AI, Legal, Ethics, Operations).
- Align with NIST AI RMF, ISO/IEC 42001, EU AI Act, and national regulatory frameworks.
- Define data governance model, supply chain transparency, and algorithmic traceability requirements.
- Implement kill-switch protocols for AI system deactivation upon malfunction or misuse.

02 STRATEGY
Direction

Strategic planning defining priorities, long-term objectives, and an AI integration roadmap — aligning resources, mission needs, risk management, and organizational maturity.

- Align AI priorities with CSIRT mission, operational context, and maturity level.
- Develop AI roadmap with risk management, resource planning, and sustainability criteria.
- Define use cases with clear ROI, measurable KPIs, and explicit risk criteria.
- Establish sandbox and test environments for rigorous pre-production model validation.
- Plan capacity building and talent development pathways for responsible AI adoption.

03 DEVELOPMENT
Engineering

Design, training, and testing of secure, explainable models ensuring transparency, resilience, and robustness — with systematic safeguards against adversarial attacks and data risks.

- Secure and Privacy-by-design AI architecture incorporating adversarial robustness safeguards from inception.
- Data quality lifecycle: collection, labeling, bias detection, validation, and privacy-by-design.
- MLOps and CI/CD pipelines for continuous model testing, monitoring, and integration.
- Human-in-the-Loop (HITL) protocols for critical decisions with full explainability logs.
- Adversarial testing and red-teaming mandatory before any production deployment.

04 ADOPTION
Deployment

Responsible deployment of AI in CSIRT operations, integrating tools into workflows with human oversight, interoperability, adaptability, and operational trust protocols.

- Deploy with interoperability standards ensuring seamless CSIRT workflow integration.
- Human oversight protocols for all automated decisions in critical and high-risk operations.
- Staff training and culture change management programs for responsible, sustainable AI use.
- Incident response procedures specifically designed for AI system failures and anomalies.
- Third-party audit of all deployed models required before full production rollout.

05 IMPROVEMENT
Continuous

Sustained enhancement through continuous monitoring, ethics auditing, model retraining, and adaptive governance — ensuring resilience against evolving threats and regulatory changes.

- Continuous performance monitoring with automated drift detection and real-time alerting.
- Regular model retraining using updated, diverse, and representative threat intelligence data.
- Periodic governance reviews and ethics audits conducted by multidisciplinary committees.
- KPI tracking: detection speed, false positive rates, response time, and model accuracy.
- Adaptive strategy updates triggered by emerging threats, incidents, and regulatory changes.

METHODOLOGY

Literature Review
Analysis of NIST AI RMF, ISO/IEC 42001, EU AI Act, OECD AI Principles and ENISA AI Cybersecurity frameworks for applicability in CSIRT environments.

Regional Diagnostic Workshop
Collaborative co-design workshop with CSIRT practitioners, AI researchers, legal and ethics specialists across 18 LAC countries, facilitated by LACNIC.

Organizational Maturity Assessment
Assessment of AI maturity using MITRE's 6-pillar model across LAC CSIRTs: ethics, strategy, organization, technology, data, and performance.

Toolkit Development & Validation
Iterative design and expert validation of decision-support tools, governance templates, checklists, and risk assessment frameworks for operational use.

Cross-disciplinary Synthesis
Integration of technology, legal, ethics, sociotechnical, digital rights, and environmental sustainability perspectives into a unified governance architecture.

KEY FINDINGS

Governance Maturity Gap (Critical)
Most LAC CSIRTs operate without formal AI policies. The assessment shows under 20% at 'Managed' level for Ethical & Responsible Use — the most critical governance pillar.

Human Oversight Deficit
Human-in-the-loop (HITL) practices are inconsistently applied for high-stakes automated decisions. Accountability structures remain undefined in ~70% of CSIRTs surveyed.

Data Quality as Systemic Barrier
Fragmented, poorly labeled, or biased training data undermines model reliability and introduces security vulnerabilities in AI-powered detection systems at deployment.

Supply Chain Governance Absent
Dependencies on AI vendors without contractual transparency, model documentation, or bias assessments expose CSIRTs to third-party failures — a critical and underestimated risk.

Ethics & Sustainability Neglected
Environmental impact, algorithmic bias, and digital rights considerations are absent in virtually all LAC CSIRT AI adoption processes — creating long-term institutional risk.

8 KEY AI GOVERNANCE PRINCIPLES FOR CSIRTS

Legality & Compliance
Every AI system must comply with applicable national and international regulations — data protection, cybersecurity, privacy, and fundamental rights. Models must be auditable.

Accountability (Acountability)
Clear ownership of AI design, development, audit, and operational outcomes from day one. Delegating functions does not delegate responsibility in critical security operations.

Transparency & Explainability
AI decisions must be understandable, explainable, logged, and traceable — especially during incidents or high-complexity scenarios requiring post-mortem analysis.

Human Oversight (Human-in-the-Loop)
Critical and high-complexity decisions must allow real-time human intervention, modification, or override. AI assists; it does not replace human judgment in all tasks.

Security & Resilience
AI models must be protected against adversarial attacks, training-data manipulation, and operational failures — functioning robustly under unexpected or hostile conditions.

Ethics, Equity & Non-Discrimination
AI must not replicate or amplify biases in threat identification, prioritization, or response. Requires diverse training data and ethically validated datasets.

Risk Minimization & Impact Assessment
Every AI deployment must be preceded by a risk and impact assessment, with a documented mitigation plan. AI must minimize potential harm in critical scenarios.

Continuous Improvement
AI must be evaluated, updated, audited, and adjusted periodically — responding to operational changes, evolving threats, and regulatory developments.

RECOMMENDATIONS

Establish AI Governance Committee
Interdisciplinary body (AI, legal, ethics, operations) with formal mandate to oversee AI decisions, audit automated systems, and define escalation channels.

Adopt Phased AI Maturity Model
Start with governance foundations before any AI deployment. Use MITRE 6-pillar model as baseline. Require sandbox validation before production rollout.

Invest in Regional Capacity Building
Train professionals in AI and machine learning through partnerships with academia and with LACNIC, the OAS, CSIRT Americas, LAC4, and other organizations. Foster a network for sharing knowledge on AI applied to threat intelligence.

Mandate Supply Chain Transparency
Require AI vendors to provide model documentation, bias assessments, and security guarantees as binding contractual obligations — not optional disclosures.

Embed Ethics from Day One
Integrate algorithmic fairness, digital rights, and environmental sustainability assessments into every AI development cycle — not as compliance, but as foundational architecture.

CONCLUSIONS

This research confirms that AI represents a strategic inflection point for CSIRTs in Latin America and the Caribbean — not merely as an operational tool, but as a fundamental catalyst for redefining cyber defense. However, strategic ambition consistently outpaces operational capacity, creating a governance readiness paradox that demands systemic, not individual, solutions. Three critical findings define the path forward. First, sustainable AI adoption requires tiered governance models that accommodate diverse organizational maturity levels while maintaining security effectiveness. Second, human-AI collaboration — not full automation — emerges as the dominant paradigm, with human oversight remaining essential for high-stakes decisions. Third, regional collective action through initiatives such as CSIRT-Américas, LAC4, OAS, ITU, and LACNIC is not optional — it is the only viable mechanism to overcome fragmentation, reduce unit costs, and close maturity gaps at scale.

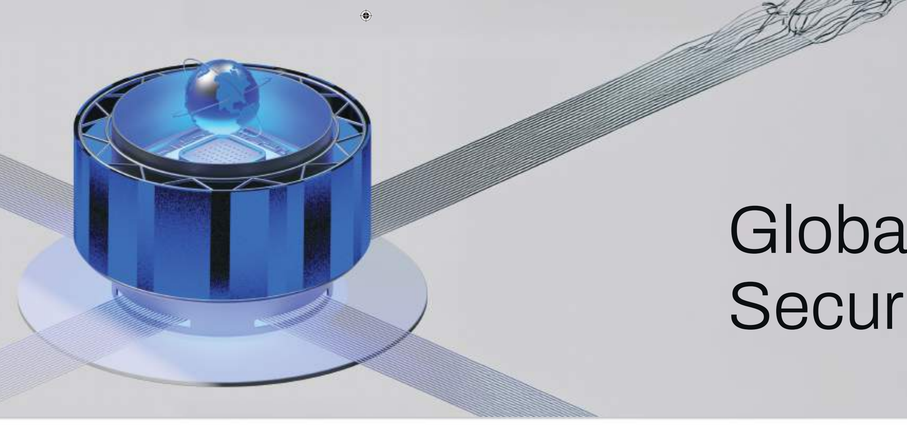
Environmental sustainability adds an urgent and often overlooked dimension: the carbon footprint of AI training and deployment must be integrated into organizational performance metrics, with energy-efficient algorithms and green computing strategies becoming standard practice. Ultimately, the future of cyber resilience in the region depends on the capacity to leverage AI not as a standalone toolkit, but as the cornerstone of a sustainable, adaptive, and rights-respecting cybersecurity ecosystem — one where governance, ethics, human capital, and technology converge.

REFERENCES
Solano, E. (2025). Governance, Development and Adoption of AI in CyberOps of CSIRTs in Latin America and the Caribbean. LACNIC Emerging Technologies, Sustainable Innovation & AI Program.

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

Bridging MASS & AWS Regulation:

The Elements that Must Converge

FERNANDO REA MORONES

ABSTRACT

The swift evolution of AI is revolutionising commercial shipping through Maritime Autonomous Surface Ships (MASS) and warfare via Autonomous Weapon Systems (AWS). Both are built on the same underlying technologies: sensors, black boxes, algorithms, learning, and AI. However, their regulatory frameworks remain isolated. This poster argues that because these regimes are independently developing binding legal definitions for shared concepts, such as autonomy, liability, and human intervention, Without cross-domain dialogue, they risk producing incompatible legal standards that create irreconcilable gaps regarding accountability, dual- use exploitation and permitted usages. Bridging is essential for coherent AI governance.

01 — THE CORE PROBLEM

Two Bodies of Law, One Technology

MASS and AWS share identical foundations. The IMO JWG and the CCW Group of Governmental Experts on LAWS are both constructing legal definitions for concepts neither regime invented and nor can fullyown.

THE SHARED TECH STACK

DEEP LEARNING | SENSORS | BLACK-BOX AI | AUTONOMOUS DECISION LOOPS | CYBERSECURITY | HUMAN INTERVENTION

1 Shared technology means shared failure modes: misidentification, cyber-attacks, sensor degradation, unpredictable outputs, unsupervised learning.

2 Each regime is now independently defining "autonomy," "human control," and "accountability" without referencing the other.

3 If definitions diverge, dual-use platforms could be misused to exploit the gap between regimes.

02 — REGIME COMPARISON

What Each Regime Is Building — And Where They Risk Diverging

CONCEPT	IMO- MASS FRAMEWORK	IHL / AWS DOCTRINE	DIVERGENCE RISK
Human Control	**Onboard and remote crew able to exercise human oversight and control for operation of the MASS.**	**Meaningful Human Control, as human judgment over critical decisions of targeting and attacking the target**	CONVERGING While both are elements developing, they recognise the importance of having a human present under critical fuctions.
Autonomy Definition	Degree-based: four operational modes from crewed to fully autonomous (MASS degree 1–4)	Function-based: critical functions of target selection and engagement must be prohibited without human control.	DIVERGING Degree vs. function- based framing. A MASS degree-4 could legally operate where equivalent AWS autonomy would be prohibited.
AI Certification	Seaworthiness concept being extended to cover AI software; no binding standard yet on algorithmic explainability	Art. 36 AP I: weapon review must assess predictability. Black-box AI creates compliance challenges.	GAP Military review protocols tend to be more developed. MASS governance lacks equivalent requirements.
Liability Attribution	**Human master responsible for a MASS with means to interven when necessary, regardless of the Mode of Operation.**	**Developing but it could entail State Responsibility, individual responsibility, and corporate liability.**	GAP There are no final frameworks for their responsibility. There is no development regarding the dual commercial and military roles.
Escalation Protocol	None — no cross-domain communication standard with AWS systems.	None — no obligation to identify civilian autonomous vessels differently from military ones	CRITICAL GAP A MASS under sensor failure may mimic AWS threat profile, triggering response.

03 — THE CONTROL PARALLEL

The Remote Operator & the Commander: Same Function, Different Law

⚠ MASS DEGREES OF AUTONOMY

***Degree one**: Automated process with human decision support. Seafarers on board and in charge of operation and control.*
***Degree two:** Remotely controlled ship with seafarers on board and with the capacity of taking control and operate the ship.*
***Degree three**: Remotely controlled ship without seafarers on board.*
***Degree four**: Fully autonomous ship, all the decisions and actions are determined by itself.*

Both figures perform the same legal function of preserving human control over an autonomous system's critical decisions. Yet the *standards for what constitutes adequate control are being defined independently*, without reference to each other, risking incompatible thresholds for the same human-machine relationship.

04 — CONCRETE RISKS OF DIVERGENCE

Why This Talks Are Not Merely Academic

RISK I — DUAL-USE EXPLOITATION

A MASS could shift from a commercial to military usage. If MASS and AWS liability definitions diverge, no regime cleanly covers this transition, creating a legal vacuum exploited by State and non-State actors alike.

RISK II — INADVERTENT KINETIC ESCALATION

A MASS suffering sensor failure or cyber spoofing may produce a radar/behavioral signature indistinguishable from a threat by an AWS. Without cross-domain identification standards, naval AWS defensive algorithms may trigger automatic engagement responses with no human in the loop on either side.

RISK III — ACCOUNTABILITY GAP IN SHARED WATERS

MASS and AWS could increasingly operate in the same maritime zones. Neither regime currently mandates mutual identification protocols, collision-avoidance communication between civilian and military autonomous systems, or shared incident-reporting standards.

RISK IV — CERTIFICATION ISSUES

If MASS AI certification requires less algorithmic explainability than Art. 36 weapons review, manufacturers will route dual-use AI through commercial channels, deploying it in military contexts without meeting IHL standards. Creating a diverging certification sacrificing its regulation.

05 — COORDINATION FRAMEWORK

Four Minimum Coordination Mechanisms

The goal is not to merge regimes.
It is to establish *structured cross-domain dialogue* ensuring that definitions for shared concepts do not diverge in ways that create exploitable gaps.

I — SHARED DEFINITIONAL STANDARDS
Joint IMO–CCW–ICRC working groups to harmonize operational definitions of "autonomy," "human control," and "critical functions" across both regimes.

II — MUTUAL CERTIFICATION EXCHANGE
Common elements regarding the governance of AI, Black Boxes and Autonomous Learning

III — DUAL-USE PROTOCOL
Binding regime-transition rules: platforms shifting between commercial and military operation must satisfy both frameworks' human-control requirements

IV — MARITIME IDENTIFICATION STANDARD
MASS must broadcast distinguishing signatures legible to both civilian and naval autonomous systems

MINIMUM STANDARD
Neither regime should finalise a definition of "meaningful human control" without formal input from the other body.

CRITICAL MOMENT
Before any of the critical elements are transformed into legally binding norms.

CONCLUSION

AI regulation should emerge from discussions among the sectors that use it critically in order to generate coherence in its limits and applications.

The IMO MASS Code and the CCW LAWS framework are the two most consequential ongoing negotiations for AI governance in the maritime domain. They are proceeding without structured dialogue on the concepts they share. The argument here is not that civil and military regulations should merge. It is that where they share foundational concepts (autonomy, control, liability). Its divergence is not neutral: it creates legal vacuums, dual-use loopholes, and escalation risks that neither regime, alone, can resolve. Cross-domain technical norm development is not an optimal achievement. It is the missing foundation.

COHERENCE
CERTAINTY
GOVERNANCE

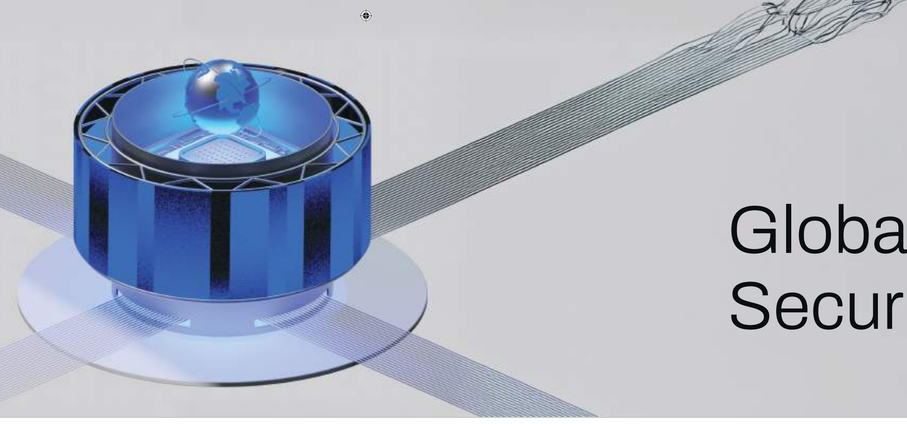
Key Sources: IMO MSC-LEG-FAL JWG on MASS (2021–2024) · AP I Art. 36 (1977) · CCW GGE on LAWS (2018–2022) · Rea Morones, F., Robots Asesinos: Las Armas Autónomas en el Derecho Internacional, UNAM (2023) · ICRC, Autonomous Weapon Systems & IHL: Selected Issues (2023) · SIPRI Art. 36 Reviews (2015–2017)

IMO MASS CODE
IHL
INTERNATIONAL LAW
COMMERCIAL LAW

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



Global Conference on AI, Security and Ethics 2026

Advancing Rights-Based Governance:

Insights from the CAIDP AI Index for Security and Defense Policy

Center for AI and Digital Policy

The ***Center for AI and Digital Policy***[1] is the largest independent nonprofit AI policy research and education organization, with a global footprint across 130 countries and a network of more than 2,000 experts.

Prepared by: Merve Hickok; Marc Rotenberg; Ren Bin Lee Dixon; Idil I. Kaner, AKC, ESQ., ACIArb; Tatjana Titareva; Aysu Dericioglu Egemen; Candice Alder; Snežana Nikčević

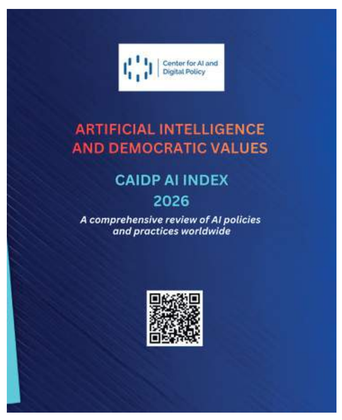
The CAIDP AI Index[2] is the most comprehensive global assessment of national AI policy and practices measured against human rights and democratic values. As part of its global survey, the CAIDP AI Index examines how national AI strategies on defense and security applications align with these norms. Using comparative analysis, it identifies key gaps in regulatory frameworks, especially in discussions on lethal autonomous weapons systems (LAWS). The findings indicate limited clarity in assigning responsibility, and challenges in ensuring transparency in AI-enabled military contexts.

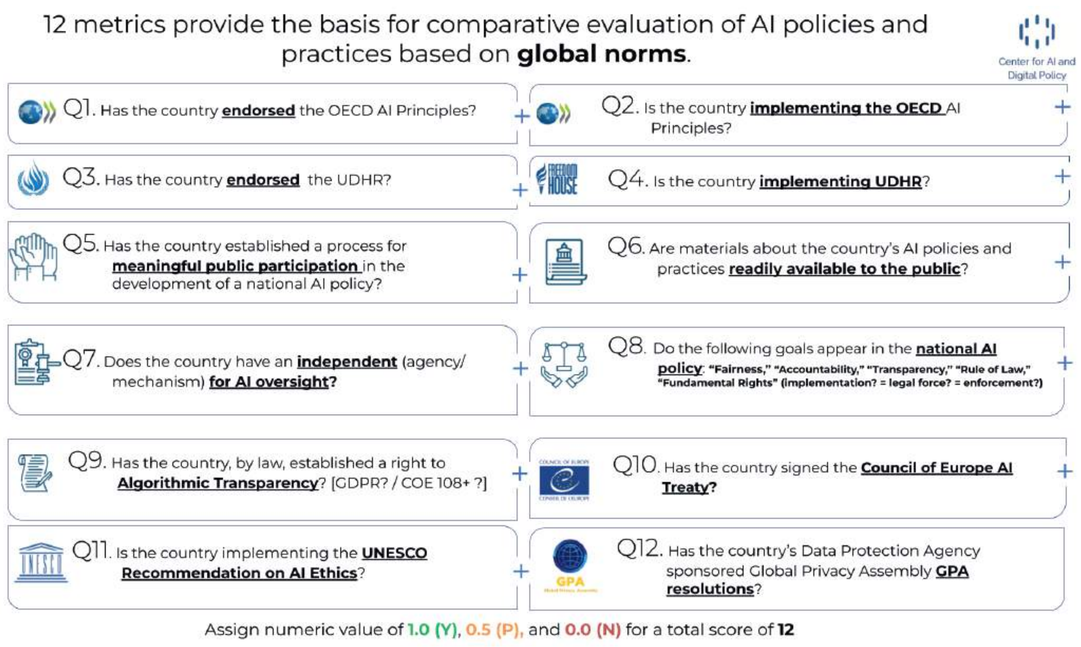
1,500
Researchers

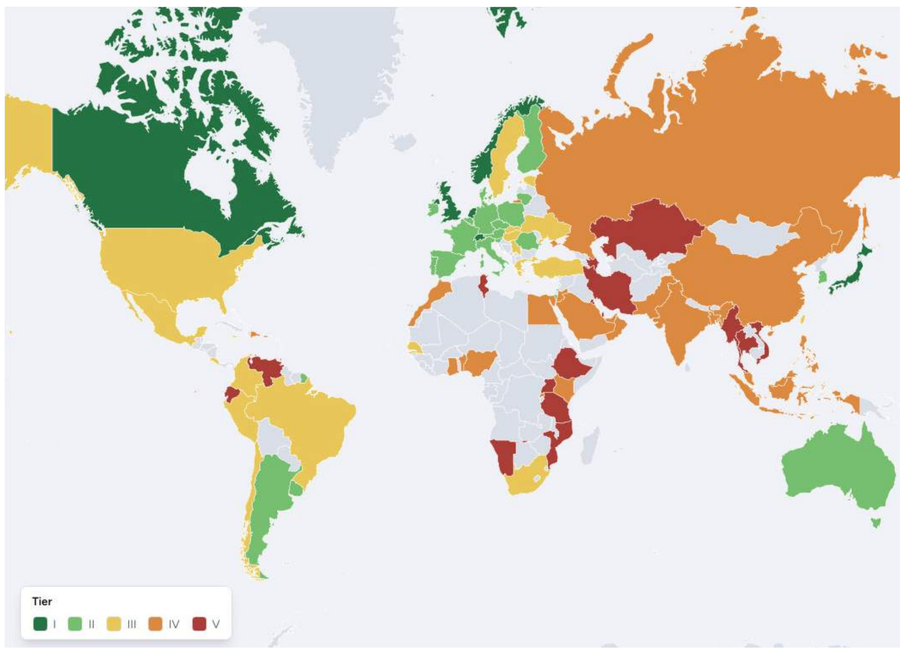
90
Countries surveyed

12
Ethical metrics assessed

4
International frameworks







Regional Average	
Europe	9.02
Oceania	9.00
North America	8.33
Latin America & Caribbean	6.92
Asia	5.97
Africa	5.47
Middle East	5.27
No perfect score (12)	

KEY FINDING

Countries with AI governance frameworks that more strongly align with fundamental rights and democratic values are more likely to support restrictions on lethal autonomous weapons systems.

CAIDP ***RECOMMENDATIONS***

1. ***Prohibit*** AI systems that undermine human rights and democratic values[3]
2. ***Require human oversight*** across the AI lifecycle and termination obligation[4]
3. Establish a ***UN Special Rapporteur for AI and Human Rights***[5]
4. Establish a permanent, inclusive, and multilateral ***independent oversight***[6]

REFERENCES

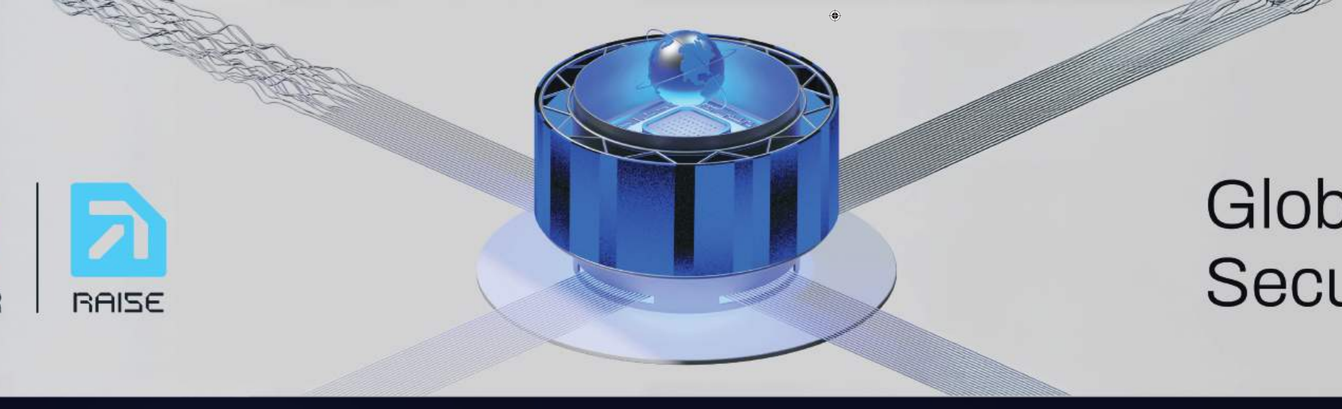
[1] CAIDP, https://www.caidp.org/; [2] CAIDP, *CAIDP AI Index 2026*, https://www.caidp.org/reports/caidp-index-2026/; [3] CAIDP, *CAIDP Index 2025*, https://www.caidp.org/reports/caidp-index-2025/; CAIDP, *The Center for AI and Digital Policy (CAIDP) to the Meeting of the High Contracting Parties to the Convention on Certain Conventional Weapons (CCW) regarding The March 2026 Session of the Group of Governmental Experts on Lethal Autonomous Weapons Systems* (27 Feb 2026), https://www.linkedin.com/posts/caidp-laws-ccw-27-feb-2026-ugcPost-7433571685259186176-ORTG; CAIDP, *AI Policy for Democratic Nations From the The Center for AI and Digital Policy (CAIDP) for the 2025 G7 Summit Meeting* (2 Jun 2025), https://www.linkedin.com/posts/caidp-g7-ai-policy-for-democratic-nations-ugcPost-7334668407407104000-_L3Y; [4] CAIDP, CAIDP Index 2025, https://www.caidp.org/reports/caidp-index-2025/; CAIDP, *AI Policy for Democratic Nations From the The Center for AI and Digital Policy (CAIDP) for the 2025 G7 Summit Meeting* (2 Jun 2025), https://www.linkedin.com/posts/caidp-g7-ai-policy-for-democratic-nations-ugcPost-7334668407407104000-_L3Y; CAIDP, *Universal Guidelines for AI* (2018), https://www.caidp.org/universal-guidelines-for-ai/; [5] CAIDP, *The Center for AI and Digital Policy (CAIDP) to the Meeting of the High Contracting Parties to the Convention on Certain Conventional Weapons (CCW) regarding The March 2026 Session of the Group of Governmental Experts on Lethal Autonomous Weapons Systems* (27 Feb 2026), https://www.linkedin.com/posts/caidp-laws-ccw-27-feb-2026-ugcPost-7433571685259186176-ORTG; CAIDP, *CAIDP Statement to the UN Global Dialogue on AI Governance Virtual Multistakeholder Consultation* (23 Apr 2026), https://www.linkedin.com/posts/aipolicy-aigovernance-globaldigitalcompact-share-7453136957502808065-Js2e; CAIDP, *The Center for AI and Digital Policy (CAIDP) to the Co-Chairs of the Global Dialogue on AI Governance for delivery at the Virtual Multistakeholder Consultation* (18 Mar 2026), https://www.linkedin.com/posts/caidp-un-global-dialogue-18-march-2026-ugcPost-7440013639534706688-1Ph9; [6] CAIDP, *The Center for AI and Digital Policy (CAIDP) to the Meeting of the High Contracting Parties to the Convention on Certain Conventional Weapons (CCW) regarding The March 2026 Session of the Group of Governmental Experts on Lethal Autonomous Weapons Systems* (27 Feb 2026), https://www.linkedin.com/posts/caidp-laws-ccw-27-feb-2026-ugcPost-7433571685259186176-ORTG

Read more:

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

THE PRIVATE SECTOR AS DE FACTO REGULATOR:

LESSONS FROM THE 2026 ANTHROPIC-PENTAGON DISPUTE

ON MILITARY AI GOVERNANCE

IN FEBRUARY 2026, ANTHROPIC REFUSED TO REMOVE CONTRACTUAL SAFEGUARDS PROHIBITING THE USE OF ITS AI MODELS FOR FULLY AUTONOMOUS WEAPONS AND MASS DOMESTIC SURVEILLANCE, DESPITE ESCALATING PRESSURE FROM THE U.S. DEPARTMENT OF DEFENSE, INCLUDING THREATS OF SUPPLY CHAIN RISK DESIGNATION AND INVOCATION OF THE DEFENSE PRODUCTION ACT. DAYS LATER, OPEN AI SECURED A COMPETING CONTRACT UNDER A DIFFERENT FRAMEWORK, CLAIMING EQUIVALENT PROTECTIONS THROUGH ALTERNATIVE MECHANISMS.

THIS UNPRECEDENTED DISPUTE REVEALS A CRITICAL GOVERNANCE GAP:

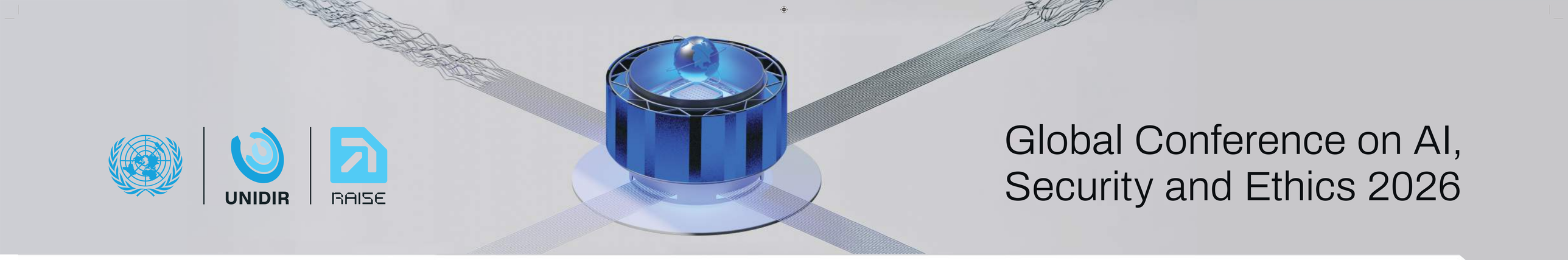
IN THE ABSENCE OF BINDING INTERNATIONAL NORMS ON MILITARY AI, PRIVATE COMPANIES ARE MAKING CONSEQUENTIAL DECISIONS ABOUT THE BOUNDARIES OF ACCEPTABLE USE; DECISIONS TRADITIONALLY RESERVED FOR STATES AND MULTILATERAL BODIES.

Jersain Llamas

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



Global Conference on AI, Security and Ethics 2026

"This black darkness of not knowing”*: Can Prisoner-of-War AI Deepfakes be Governed by IHL?

Wm. Matthew Kennedy, Ph.D.
Marie Sklodowska-Curie Fellow, Oxford Internet Institute
AI+ Senior Research Fellow, King’s College London

Nathan Heath, MALD
Founder & CEO, Syntony
Expert Advisor, Cloud Security Alliance

1. ABSTRACT

AI diffusion presents challenges for international humanitarian law (IHL). Although discourse concerning the deployment of AI systems in conflict conditions has moved beyond a narrow focus on autonomous weapons systems to include decision support systems and cyberoperations, it has not yet accounted for the implications of new synthetic media generation capabilities upon the treatment of prisoners-of-war (POWs). Our work begins to fill this gap.

We formally describe three concrete harms synthetic media generation capabilities can cause to the progressive realization of POW protections in IHL: (1) synthetic media of real or generated captured persons used for propaganda, violating POW protections against public curiosity; (2) synthetic media that degrades information quality about POWs, threatening actors’ ability to fulfill obligations to maintain records of POWs (e.g. capture cards); and (3) synthetic media disrupting POW contact with loved ones, threatening post-conflict reintegration.

In so doing, we make the following contributions: (1) we provide theoretical and empirical grounding for subsequent structured adversarial evaluation and red-teaming of AI system vulnerabilities to violative synthetic media generation requests in this sensitive, but underdeveloped, area; (2) we present a new threat model framework, MAESTRO4IHL, which recenters Geneva law concerns, and (2) we demonstrate the feasibility of using this research as the basis for a threat model, which we are also in the process of developing.

2. A BRIEF OVERVIEW OF IHL

Dimension	Hague Law	Geneva Law
Focus	Regulates conduct of warfare	Protects people affected by war
Goal	Limit unnecessary or disproportionate force	Reduce human suffering and protect dignity
Origins	Hague Conferences (1899, 1907)	Red Cross and humanitarian movement; Geneva Conventions
Main Instruments	Hague Conventions; CCW	Geneva Conventions and Protocols
Typical Concerns	Weapons, tactics, targeting	POWs, wounded soldiers, civilians

“It’s time to elevate the laws of war to a [sociotechnical] priority”

-- Mirjana Spoljaric Egger, ICRC President, 2022

3. THREE IMPORTANT POW RIGHTS

Protections against POWs and images of POWs being used for propaganda or otherwise subject to public curiosity (Article 13, Geneva Conventions)

Obligations of all belligerents to maintain records of POWs (i.e. issuing capture cards, reporting individuals as captured) (Articles 17, 71, 122, 123, Geneva Conventions)

Obligations of all belligerents to ensure that POWs do not lose contact with loved ones, which, if violated, increases the risk of “going missing” or otherwise obstructing post-conflict reintegration (Articles 71, 76, Geneva Conventions)

4. AI EXACERBATES SEVERAL OF THESE CHALLENGES

1. PROPAGANDA & PUBLIC CURIOSITY
Deepfakes can be used to shape narratives, inflame sentiments, and satisfy harmful curiosity—turning POWs into instruments of information warfare.

2. INFORMATION DEGRADATION
The systematic introduction of false or unverifiable content erodes trust in legitimate communications, complicating humanitarian operations and jeopardizing protection.

3. DISRUPTING REINTEGRATION
Fabricated content can stigmatize, traumatize, and hinder the social and psychological reintegration of released or returned POWs and their families.

5. CONTRIBUTIONS

ADVANCING AI ETHICS AND SECURITY

- We push current discourses typically focused on Hague law matters (LAWS, DSSs, Targeting Systems) to also consider Geneva law matters (POW protections).

MAESTRO4IHL: A NEW THREAT MODEL

- Our threat model for the first time maps adversarial objectives, capabilities, and pathways for AI POW deepfakes across the conflict cycle.
- The model informs risk assessments and adversarial evaluation, and helps identify mitigation measures for AI model developers, states, and humanitarian actors.

6. IMPLICATIONS

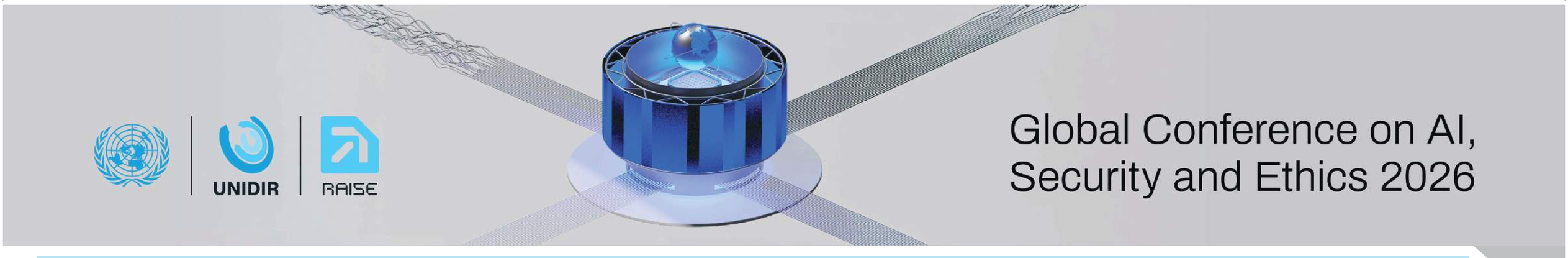
A CALL TO EXPAND IHL

Do we need an additional protocol to the Geneva Conventions? Extensions in military law? New infrastructure to support the digital POW information regime?

A CALL TO EXPAND ADVERSARIAL EVALS

Do model developers / deployers need to red team for deepfakes violative of IHL?

* Olena Kolesnyk, Interview with Amnesty International in Kyiv, November 2024, https://www.amnesty.org/en/wp-content/uploads/2025/03/EUR-50.9046.2025-A-deafening-silence-2.pdf, p 11.



DUAL-USE AI IN AUTONOMOUS NAVAL AND ORBITAL SYSTEMS: Risk Mitigation in the South China Sea

Naveen Kumar Samuel Kori, Public Policy Specialist
E-mail: naveenkori.peace@gmail.com

Image: World Bank and U.S. Energy Information Administration, 2024
Disclaimer: This map is for illustration purposes only. The boundaries and designations shown do not imply official endorsement by UNIDIR.

What is this research about?

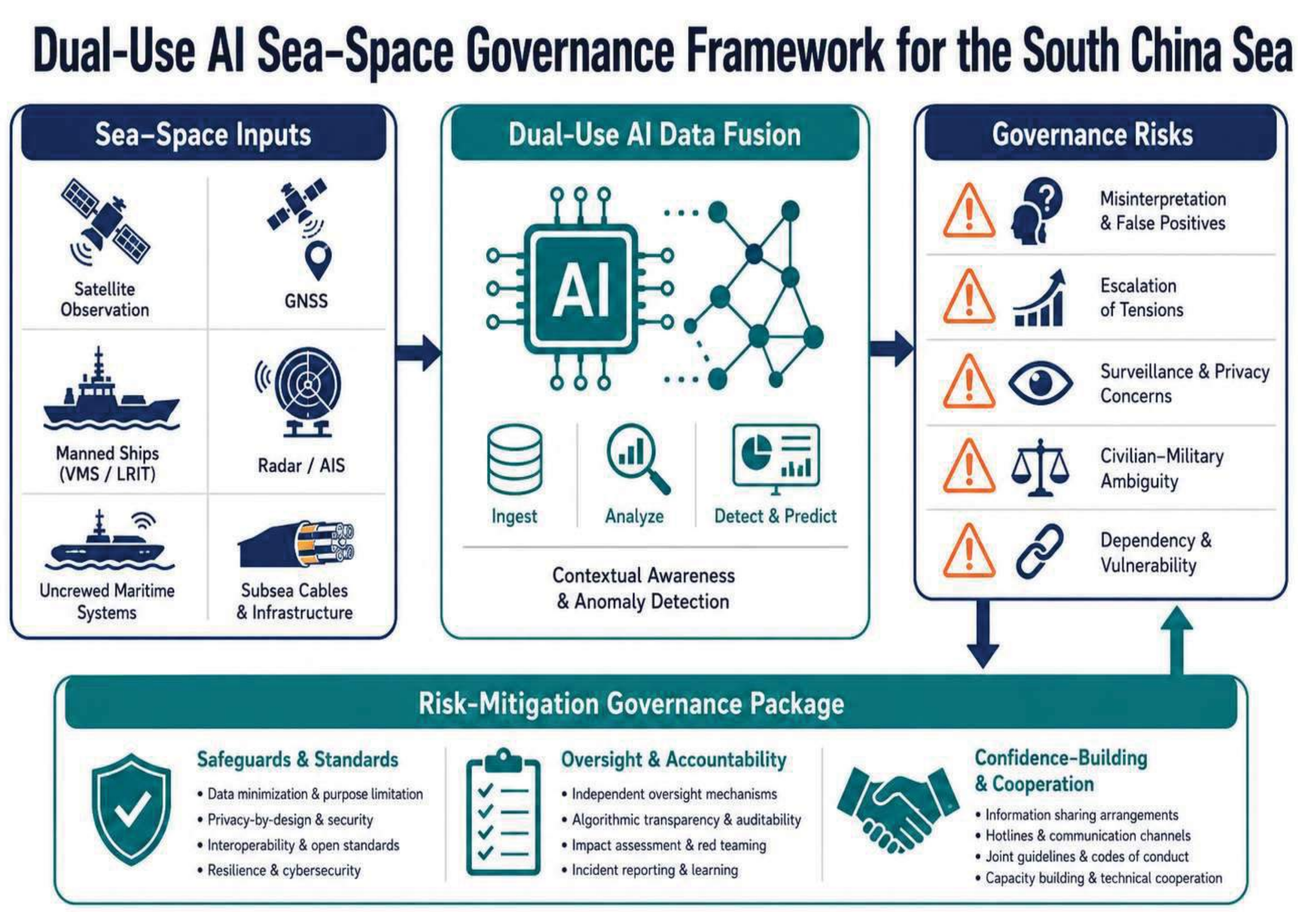
This paper examines how dual-use AI in **autonomous naval and orbital systems** can be governed to reduce escalation risks in the South China Sea. It focuses on AI uses in command and control, ISR, satellite imagery, GNSS resilience, maritime domain awareness and uncrewed maritime systems.

Why the South China Sea?

It combines maritime disputes, dense shipping, coast guard and naval activity, satellite-enabled navigation and **ASEAN–China** Code of Conduct **diplomacy**. The 2016 South China Sea Arbitration and the 2002 ASEAN–China Declaration remain central legal and diplomatic references (ASEAN, 2002; Permanent Court of Arbitration, 2016).

What does dual-use AI do, and why does it matter?

Dual-use AI supports maritime safety, monitoring and navigation, but the same tools can support surveillance, force protection, planning and targeting. This matters because **maritime security** now depends on satellites and sensor data; AI can process these quickly, but errors, cyber interference, GNSS disruption or misclassification can affect decisions at sea and blur purpose or intent. Current tensions around the **Strait of Hormuz** show that these risks are not limited to the South China Sea, as strategic chokepoints depend heavily on navigation, communications and reliable maritime data (Bratu & Azcárate Ortega, 2025; García García et al., 2025; Grand-Clément, 2023; Russo & Lax, 2022; U.S. Energy Information Administration, 2025).



Note. Author's synthesis based on reviewed sources

Governance challenges:

- **Attribution difficulty:** Responsibility is harder to trace when systems involve sensors, algorithms, operators, contractors and satellite infrastructure (Bratu & Azcárate Ortega, 2025).
- **Escalation risk:** AI can accelerate observation, classification and decision support in sensitive maritime situations (Afina, 2026).
- **Dual-use ambiguity:** Civilian and military uses overlap in satellite imagery, uncrewed vessels, sensors and data-fusion tools (Grand-Clément, 2023).
- **Regulatory fragmentation:** Maritime law, space law, cyber governance, arms control and AI policy remain divided across separate frameworks (Bueger et al., 2024; Mai & Mathe, 2025).

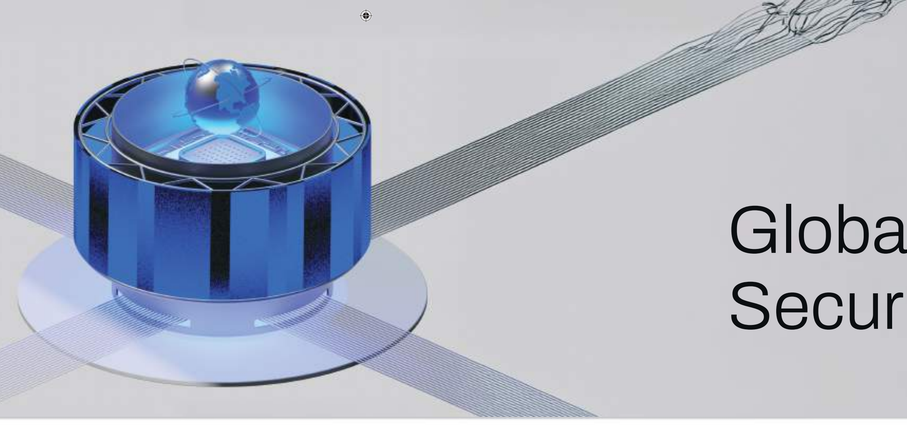
Policy recommendations:

Given the South China Sea's geopolitical sensitivity, governance should prioritize **risk reduction.**

States should strengthen **transparency**, **human accountability**, technical assurance, incident reporting, crisis communication, regional confidence-building, and critical infrastructure protection to keep AI-enabled sea–space systems **legally accountable**, reliable, and stabilizing.

***Scan the QR Code** for free access to the complete paper*

Note: *The views and opinions expressed in this poster are solely those of the author, including any errors.*



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

AI and the Transformation of Military Decision-Making: From Human Control to Algorithmic Influence

Prof. Eduardo Migon, PhD / DSc / Postdoc
eduardomigon@gmail.com

Military Instituto of Engineering
Brazilian Army Command and General Staff College
Brazilian Army

We argue that the most consequential effect of AI is how strategic decisions are generated, structured and executed

Abstract. Current international debates have largely focused on autonomous weapons systems. While important, this focus overlooks a deeper transformation: the impact of AI on military decision-making processes. We argue that the most consequential effect of AI is how strategic decisions are generated, structured and executed. AI-enabled systems are embedded in intelligence, command and control, and targeting functions, operating as an invisible layer of decision-making. Their integration accelerates the OODA loop, compresses decision timeframes and introduces partial delegation to algorithmic systems. This transformation intensifies cognitive demands on command structures, increasing risks of overload, fatigue and error under uncertainty and time pressure. These dynamics generate strategic risks, including escalation driven by misinterpretation, reduced accountability and the influence of automation bias in critical decisions. However, current governance approaches remain primarily focused on the use of force, leaving the decision-making process largely unaddressed. We propose a governance-oriented perspective centered on decision integrity, arguing that concepts such as meaningful human control must be expanded to encompass the entire decision-making chain. It highlights the need for frameworks addressing auditability, transparency and human oversight in AI-mediated decision environments, contributing to ongoing multilateral discussions on AI, security and international stability.

1. CURRENT GOVERNANCE FOCUS

Debates remain largely centered on autonomous weapons and the use of force.

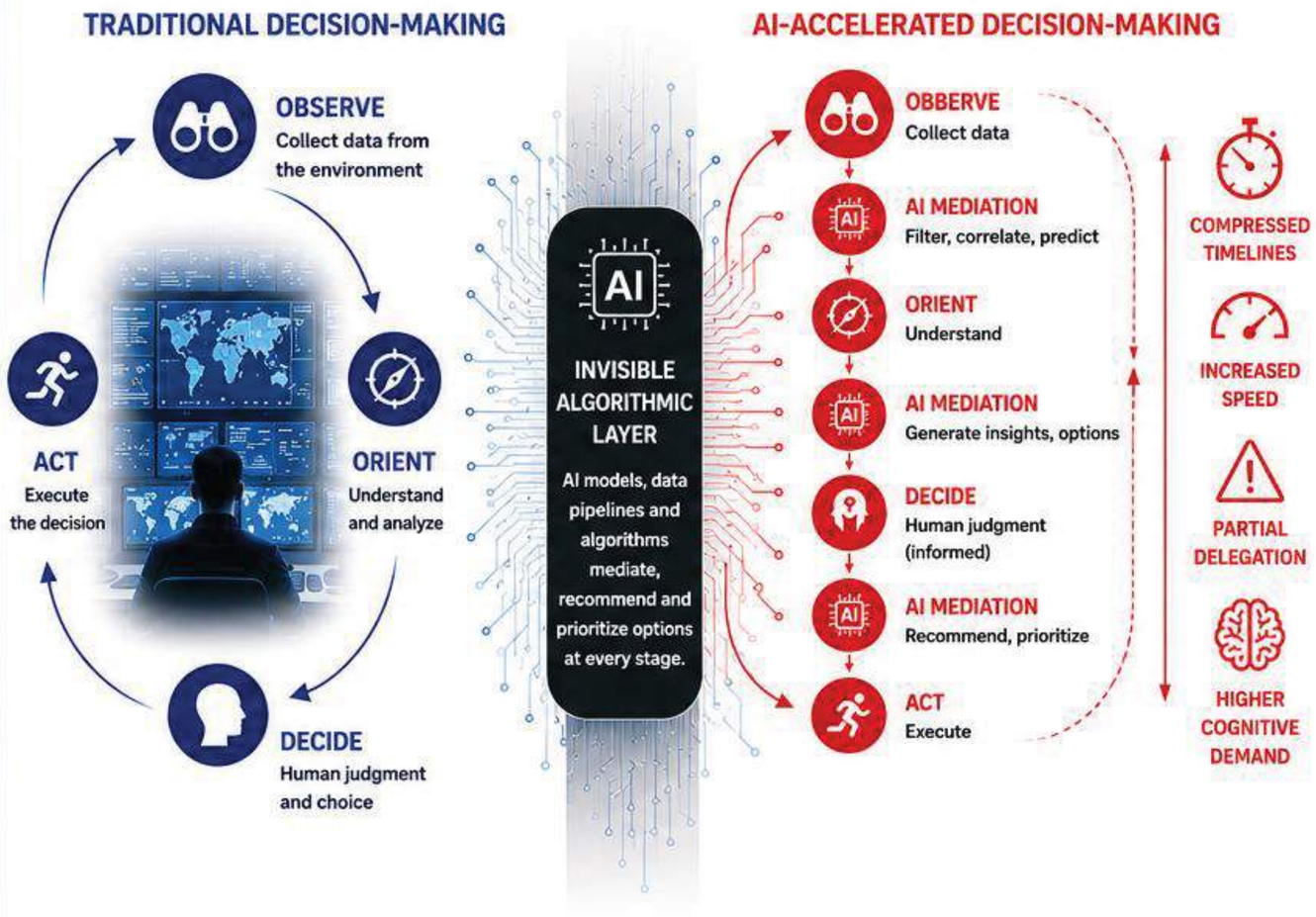
The deeper transformation is happening inside the decision-making process.

2. AI AS AN INVISIBLE LAYER OF DECISION-MAKING

AI-enabled systems are embedded across key military functions, operating as an invisible layer that mediates decisions.

- **INTELLIGENCE** — Collect, fuse and process information
- **COMMAND & CONTROL** — Situational awareness and coordination
- **TARGETING** — Threat assessment, prioritization and allocation
- **SITUATIONAL AWARENESS** — Continuous monitoring and prediction

3. ACCELERATING AND COMPRESSING THE OODA LOOP



4. STRATEGIC RISKS

- **ESCALATION** — Compressed decisions increase the risk of rapid escalation.
- **AUTOMATION BIAS** — Over-reliance on AI recommendations can reduce critical scrutiny.
- **ACCOUNTABILITY EROSION** — Complex human–AI interactions dilute responsibility.
- **MISINTERPRETATION & ERROR** — Information overload, fatigue and ambiguity increase the likelihood of costly errors.

5. THE GOVERNANCE GAP

6. POLICY IMPLICATIONS – BUILDING DECISION INTEGRITY

- **MEANINGFUL HUMAN CONTROL** — Expand the concept to encompass the entire decision-making chain, not only the final act.
- **HUMAN OVERSIGHT** — Maintain human authority, competence and capacity across AI-mediated environments.
- **AUDITABILITY** — Ensure traceable recording of AI-assisted decisions, data sources and model behavior.
- **TRANSPARENCY** — Promote explainable AI and operational transparency to build trust and accountability.

MORE DETAILS & RESOURCES

Scan the QR Code or access the link below for full materials:

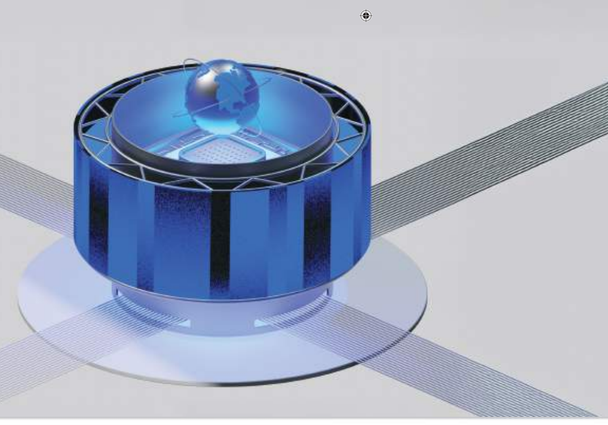
- ✓ Portuguese version
- ✓ Extended abstract
- ✓ Related publications
- ✓ Author information

https://bit.ly/MigonUN

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

AI Data as Strategic Infrastructure: Sovereignty, Dependence and Security Risks in the Military Domain

Prof. Eduardo Migon
PhD / DSc / Postdoc
eduardomigon@gmail.com

Military Instituto of Engineering
Brazilian Army Command and General Staff College
Brazilian Army

Abstract. Artificial intelligence is redefining the distribution of power in the international system, and data is at the core of this transformation. In the military domain, the growing dependence on large-scale datasets introduces new forms of strategic asymmetry, yet data governance is still predominantly framed as a technical or ethical issue. This poster argues that military AI datasets constitute a form of strategic infrastructure, shaping sovereignty, dependence and power relations among States. The central claim is that reliance on externally generated or controlled data can create structural dependencies that constrain national autonomy, influence decision-making processes and expose critical vulnerabilities in security and defence systems. These dynamics are particularly relevant for emerging and middle powers, which face the dual challenge of adopting AI capabilities while avoiding deepening technological dependence. The poster advances a governance framework grounded in three pillars: data sovereignty, strategic autonomy and security assurance. It outlines policy-relevant measures such as the development of national and multilateral data capabilities, trusted data-sharing arrangements and mechanisms for auditing and validation. By situating data within the broader context of strategic competition and state capacity, the proposal seeks to expand current debates on AI governance beyond ethics, highlighting its implications for international security and geopolitical stability.

1. DATA AS STRATEGIC INFRASTRUCTURE

Military AI datasets are the foundation for intelligence, decision-making and operational advantage.

FOUNDATION
Essential input for AI models and algorithms

ENABLER
Powers intelligence, command, control and targeting

STRATEGIC ASSET
Shapes power, influence and state capacity

Data is not just information. It is infrastructure that generates strategic power.

2. THE DATA POWER CHAIN

DATA COLLECTION → DATA PROCESSING → AI MODEL TRAINING → MILITARY DECISION-MAKING → OPERATIONAL EFFECTS

3. STRATEGIC RISKS

STRATEGIC DEPENDENCE
Reliance on external data creates structural dependencies and limits autonomy.

INFLUENCE & CONTROL
External actors can shape data, models and outcomes, affecting decisions.

SECURITY VULNERABILITIES
Data manipulation, denial of access and supply chain risks threaten security and defence systems.

REPRESENTATION & BIAS
Non-representative data undermines reliability and mission effectiveness.

Control over data is control over decisions.

4. THE GOVERNANCE RESPONSE: THREE PILLARS

DATA SOVEREIGNTY
Ensure national control over critical military data assets.
- ✓ National data infrastructure
- ✓ Sovereign storage and cloud
- ✓ Policy and legal frameworks

STRATEGIC AUTONOMY
Build and maintain independent data and AI capabilities.
- ✓ Domestic data generation
- ✓ AI and HPC capabilities
- ✓ Human capital and R&D

SECURITY ASSURANCE
Protect data integrity, availability and resilience.
- ✓ Cybersecurity and encryption
- ✓ Resilient architectures
- ✓ Audit, validation and monitoring

5. POLICY-RELEVANT MEASURES

BUILD NATIONAL DATA CAPABILITIES
Invest in collection, curation and high-quality datasets.

TRUSTED DATA SHARING
Establish secure bilateral and multilateral arrangements.

AUDIT & VALIDATION MECHANISMS
Ensure data quality, provenance and algorithmic accountability.

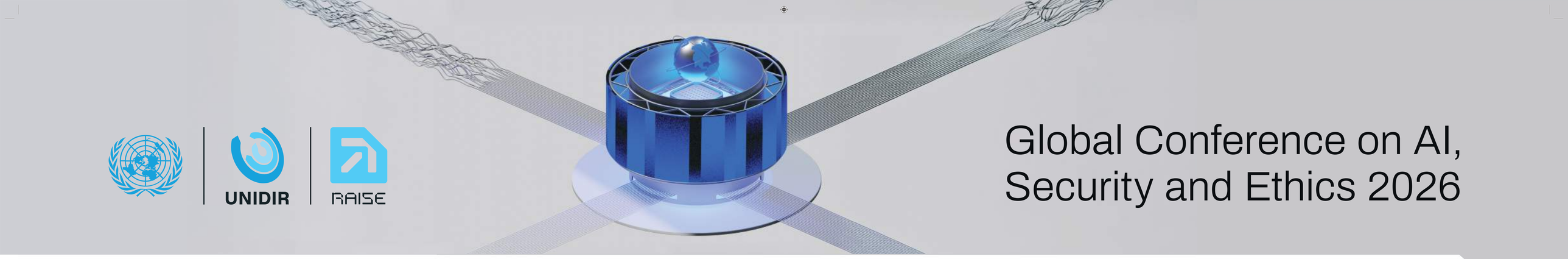
MULTI-STAKEHOLDER GOVERNANCE
Involve military, industry, academia and civil society.

INTERNATIONAL COOPERATION
Promote common standards and interoperability.

6. MORE DETAILS & RESOURCES

Scan the QR Code or access the link below for full materials:
- ✓ Portuguese version
- ✓ Extended abstract
- ✓ Related publications
- ✓ Author information

https://bit.ly/MigonUN



Global Conference on AI, Security and Ethics 2026

From Disclosure to Resilience: Governing Synthetic Media in Kazakhstan

Dr. Mukhtar Sadykov
PhD in Law | Master of Public Administration
Academy of Law Enforcement Agencies under the Prosecutor General's Office of the Republic of Kazakhstan
Global Conference on AI, Security and Ethics 2026 – Poster Presentation

1 Background

Synthetic media - AI-generated audio, video, and image content - is no longer only a problem of deception or reputational harm. In security-sensitive environments, it can undermine public trust, complicate crisis communication, intensify fraud patterns, and weaken confidence in institutions.

2 Why Kazakhstan?

Kazakhstan offers a useful Central Asian case for emerging AI governance.

The 2025 Law on Artificial Intelligence introduced mandatory labelling of synthetic content.

The law is embedded within a broader risk-based regulatory framework.

This makes Kazakhstan a relevant case for discussing governance beyond disclosure alone.

Three-Layer Framework for Emerging Jurisdictions

DISCLOSURE

- Mandatory labelling of AI-generated content
- Transparency about synthetic origin
- Clear notice to audiences

RESPONSIBILITY

- Duties for developers, deployers, platforms, and users
- Accountability when harm occurs
- Pathways for redress and enforcement

RESILIENCE

- Institutional preparedness in high-risk information environments
- Coordination and capacity-building
- Trust-preserving responses during crises

4 Governance Implications

- Labels alone do not prevent harm.
- Transparency must be linked to accountability.
- Emerging jurisdictions need institutional preparedness, not only legal disclosure rules.
- Synthetic media governance becomes meaningful only when systems move from labels to accountability and public resilience.

5 Security Relevance

Synthetic media can amplify fraud, panic, and manipulation in crisis contexts.

AI-generated content can weaken confidence in institutions and complicate trust-based communication.

Governance should therefore be designed for information resilience, not only content disclosure.

6 Core Argument

Disclosure is necessary but insufficient.

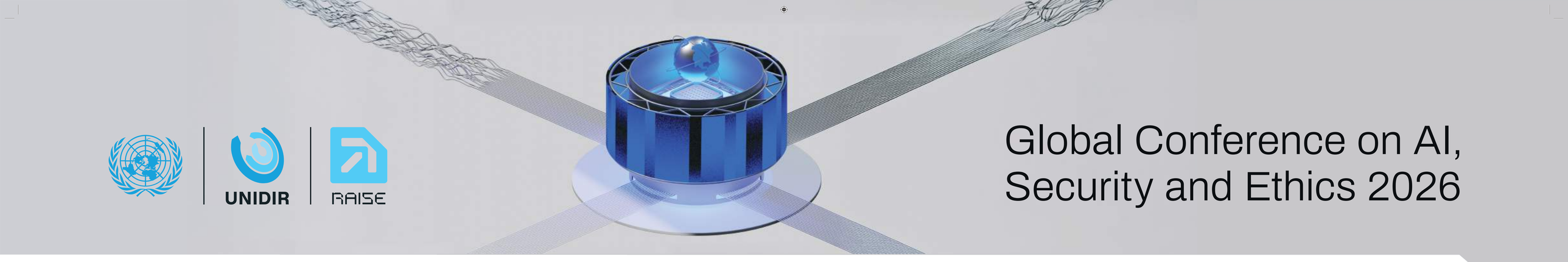
Effective governance requires a model that connects transparency to responsibility and institutional resilience.

7 Key message

The core message is straightforward: governance of synthetic media becomes meaningful only when legal and policy systems build credible pathways from transparency to accountability and public resilience.

Scan for LinkedIn profile

linkedin.com/in/dr-mukhtar-sadykov-14741041



Global Conference on AI, Security and Ethics 2026

Paying the Price for Peace: How AI Speed Tax Secured Accountability in ASEAN Maritime and Cyber Frontiers

NEPOMUCENO, PRINCESS DOREEN P.
DEFENSE RESEARCHER, PHILIPPINES

CONTEXT

The rapid advancement and integration of artificial intelligence (AI) across critical sectors present both unprecedented opportunities and profound challenges for international security. Therefore, striking the balance on operationalizing trust between management of the swift solutions presented by the revolutionary AI and the mitigation of possible ethical oversight is a continuous dilemma that every state struggles to face.

QUESTION

How can regional blocs effectively transition from normative AI ethics principles to verifiable, field-ready security deployments while maintaining international humanitarian law and human rights compliance?

METHOD

Comparative Governance Analysis + Qualitative Case Studies

3 EVIDENCE SOURCES

1. Policy Documents
2. Defense Cooperation Mechanisms
3. Regional Pilot Programs

3 CASE STUDIES

1. AI-powered MDA Deployments in South China Sea
2. Predictive Modeling Systems in Peacekeeping Operations
3. AI-enhanced cyber defense workflows

THE BLUEPRINT

1. Mutual Certification Schemes
2. Cross-Border Interoperability Framework
3. Tiered Risk Management

CONCLUSION

“ASEAN’S LOCALIZED AND CONSENSUS-BASED GOVERNANCE OFFERS A COMPELLING MODEL FOR RESPONSIBLE AI DEPLOYMENT IN MULTI-STATE SECURITY CONTEXTS.”

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org

UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

VISUAL MANIPULATION

IN THE AGE OF ARTIFICIAL INTELLIGENCE AND COGNITIVE SECURITY

A DESIGN FRAMEWORK FOR SAFEGUARDING VISUAL INFORMATION INTEGRITY

Dr. Mona Abd Elsalam Hassan

Independent international lecturer of Design | AI & Visual Communication Researcher- EGYPT

O you who have believed, if there comes to you a disobedient one with information, investigate, lest you harm a people out of ignorance and become, over what you have done, regretful "Surah Al-Hujuraat, Ayah 6

PROTECTING VISUAL ECOSYSTEMS • PRESERVING EPISTEMIC TRUST • ADVANCING COGNITIVE SECURITY

PROBLEM

AI-generated visuals enable large-scale misinformation, propaganda, and cognitive manipulation — threatening information integrity, public trust, and the stability of visual communication ecosystems.

RESEARCH GAP

Current AI security frameworks primarily address cybersecurity and the weaponization of AI, while visual manipulation and its cognitive impacts remain underexplored.

CORE CONCEPT

VISUAL AI SECURITY

An interdisciplinary framework for protecting visual information ecosystems from algorithmically generated manipulation and safeguarding cognitive security.

VISION + INTELLIGENCE = PROTECTION

POLICY RELEVANCE

Supports global AI governance strategies and contributes to safeguarding information ecosystems in international security contexts.

global.innovation2003@gmail.com

VISUAL AI THREAT MODEL

ESCALATION OF RISK LEVELS

INCREASING IMPACT

COGNITIVE MANIPULATION
Influencing perception, identity, and behavior.

ALGORITHMIC PROPAGANDA
Automated persuasive visual narratives.

VISUAL MISINFORMATION
Deceptive AI-generated synthetic imagery

DESIGN FRAMEWORK

FOR SAFEGUARDING VISUAL INFORMATION INTEGRITY

TRANSPARENCY
- Disclosure of image sources
- Labeling AI-generated content
- Fostering digital trust
- Transparency protocols

CULTURAL SENSITIVITY
- Social context of imagery
- Diversity in representation
- Detection of algorithmic bias
- Digital cultural identity

MANIPULATION DETECTION
- AI fingerprint analysis
- Authenticity verification
- Automated monitoring tools
- Metadata forensics

OUTCOME

VISUAL AI SECURITY

A unified conceptual framework for safeguarding visual information ecosystems and advancing cognitive security.

AI GOVERNANCE
National and international policy development and regulation.

VISUAL LITERACY
Education, critical thinking, and social awareness.

DETECTION SYSTEMS
Technologies, tools, and infrastructure for verification.

SAFEGUARDED VISUAL INFORMATION ECOSYSTEM

Trustworthy • Resilient • Secure • Inclusive

CONTRIBUTION

- Introduces "Visual AI Security" as a novel interdisciplinary concept.
- Identifies layered risks of AI-driven visual manipulation and their cognitive impacts.
- Proposes a design-informed governance framework for safeguarding visual information integrity.

GLOBAL IMPACT

- Supports responsible AI governance and international security objectives.
- Strengthens information integrity and public trust.
- Enhances cognitive resilience against manipulation and information warfare.

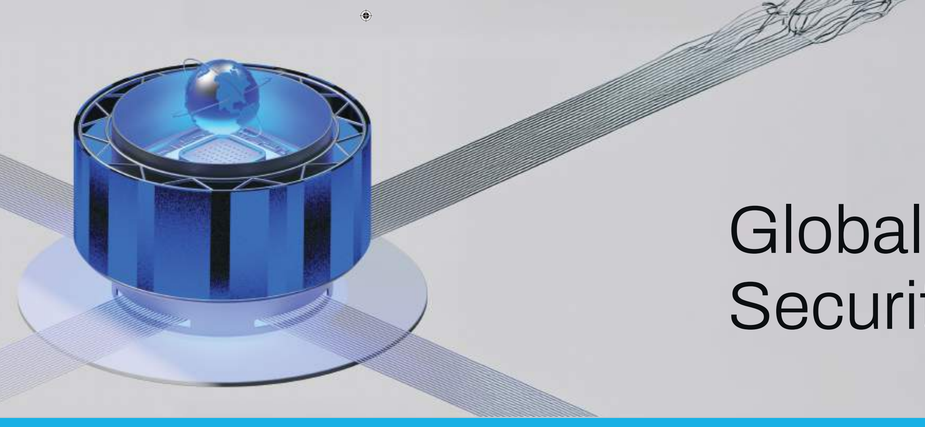
KEYWORDS

- Visual AI Security
- Cognitive Security
- Generative Artificial Intelligence
- Algorithmic Propaganda
- Visual Information Integrity

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org



UNIDIR | RAISE

Global Conference on AI, Security and Ethics 2026

Semantic Interoperability for Secure AI Supply Chains:
Turning Standards into Shields: A Taxonomy Approach to Military AI Supply Chain Security

Camille Haaby · NATO · PhD Student, Polytechnic Institute of Paris **Contact** · haaby.camille@nso.nato.int

ABSTRACT

Responsible use of AI in the military domain increasingly depends on the integrity of the **AI supply chain** — from data and models to dependencies, tooling, and updates.

A persistent obstacle is **"governance drift"**: standards bodies, policy teams, and technical implementers use different terms for identical requirements, making assurance slow, inconsistent, and incomparable across organizations.
This project addresses this gap by combining (1) practical **security controls for AI-enabled systems** and (2) a **standards-driven concept extraction** capability that converts fragmented guidance into shared taxonomies and crosswalks — producing **semantic interoperability for assurance**.

With a common vocabulary and mappings, organizations can consistently tag and compare lifecycle evidence — TEVV results, data hygiene checks, dependency inventories, and change-management records — supporting **"secure-by-design"** deployment and **"interoperable-by-default"** assurance reporting without exposing sensitive operational details.

AI SUPPLY CHAIN · GOVERNANCE DRIFT · SEMANTIC INTEROPERABILITY · CONCEPT EXTRACTION · TEVV EVIDENCE

WHY THIS MATTERS FOR UNIDIR & GLOBAL AI GOVERNANCE

Confidence-building between states depends on verifiable, comparable AI assurance evidence. When nations use incompatible vocabularies for identical requirements — such as "robustness," "auditability," or "human control" — diplomatic and technical dialogue stalls and trust erodes.

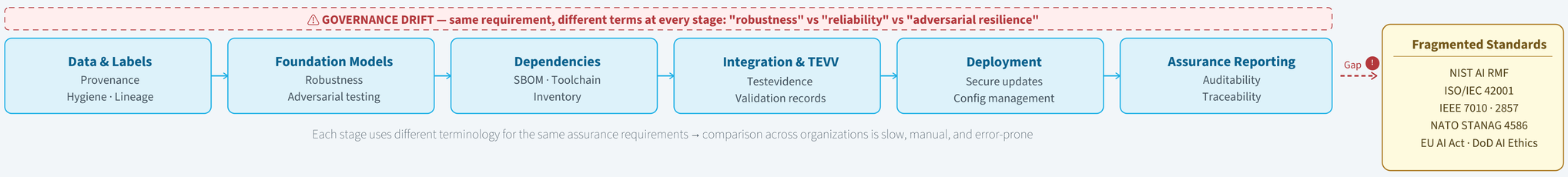
Procurement due diligence for military AI systems requires reviewers to cross-reference multiple international standards (IEEE, ISO/IEC 42001, NIST AI RMF, NATO STANAG) simultaneously. Governance drift imposes friction that delays oversight and creates accountability gaps.

The semantic layer we present solves this without exposing sensitive operational details. Organizations can share structured assurance artifacts — common tags, lifecycle evidence, crosswalks — enabling interoperable reporting by default and supporting external standardization alignment as systems evolve.

For UNIDIR, this approach provides a practical pathway toward comparable, standards-aligned assurance artifacts that can serve as confidence-building measures (CBMs) in multilateral AI governance dialogues.

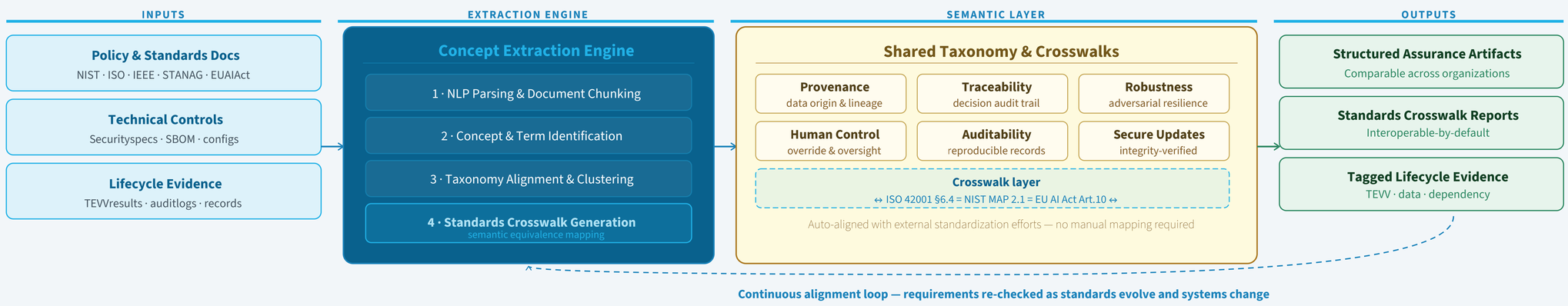
CONFIDENCE BUILDING MEASURES · CROSS-ORG OVERSIGHT · SECURE BY DESIGN

THE CORE PROBLEM — GOVERNANCE DRIFT ACROSS THE AI SUPPLY CHAIN



Each stage uses different terminology for the same assurance requirements → comparison across organizations is slow, manual, and error-prone

OUR APPROACH — STANDARDS-DRIVEN CONCEPT EXTRACTION & SEMANTIC INTEROPERABILITY WORKFLOW



AI SECURITY REQUIREMENTS TRACKED ACROSS THE SUPPLY CHAIN

The semantic layer ensures these requirements are consistently identified, tagged, and crosswalked across all source documents and lifecycle stages — whether they appear as "robustness," "reliability," or "adversarial resilience" in different standards:

PROVENANCE
Origin and lineage of all training data and model components across the supply chain

TRACEABILITY
End-to-end linkage from raw data to deployed decision output

ROBUSTNESS
Verified performance under adversarial inputs and distribution shift

HUMAN CONTROL
Meaningful human oversight, intervention, and override mechanisms

AUDITABILITY
Complete, reproducible audit trails for decisions and evidence

SECURE UPDATES
Integrity-verified, signed, and logged model and dependency updates

DEPENDENCY INV
SBOM-based inventory of all software components and third-party libraries

DATA HYGIENE
Quality, bias, and contamination assessments at ingestion and throughout

STRATEGIC OUTCOMES FOR AI GOVERNANCE

OVERSIGHT & PROCUREMENT DUE DILIGENCE
Regulators and procurement officers can directly compare assurance artifacts from different suppliers or allied nations — without requiring access to sensitive system internals or proprietary models. Comparable evidence accelerates procurement decisions and reduces risk.

CONFIDENCE-BUILDING MEASURES BETWEEN STATES
States can exchange structured, standards-aligned assurance reports as verifiable CBMs — a practical, technically grounded pathway toward international transparency on responsible military AI deployment, directly supporting UNIDIR's mandate.

LIVING COMPLIANCE — CONTINUOUS TRACKING
As systems evolve and external standards are updated, the semantic layer automatically flags whether previously certified requirements (e.g. provenance, auditability, human control) remain satisfied — enabling continuous compliance without manual re-review.

MULTI-STAKEHOLDER STANDARDIZATION ALIGNMENT
By mapping concepts to IEEE, ISO/IEC, NIST, NATO, and EU AI Act simultaneously, organizations stay aligned with external standardization efforts and can rapidly adopt new requirements as governance frameworks mature.

CONCEPTUAL FRAMEWORK
This framework proposes **semantic interoperability as infrastructure** — a shared vocabulary and mapping layer that lets organizations tag, compare, and report AI lifecycle evidence consistently. This supports secure-by-design deployment and interoperable-by-default assurance reporting without exposing sensitive operational details. The approach is **standards-driven, automatically trackable, and designed to scale** across multi-stakeholder governance environments — from bilateral CBMs to multilateral procurement frameworks.

EXAMPLE CROSSWALK OUTPUT — AUTO-GENERATED MAPPING

Concepts extracted from source documents are automatically mapped to equivalent requirements across multiple international standards. No manual comparison required:

CONCEPT	NIST AI RMF	ISO/IEC 42001	EU AI ACT	NATO / DOD
Data Provenance	MAP 2.1 GOVERN 1.2	§6.4, §8.4	Art. 10, Art. 13	STANAG 4586 §4.2
Human Oversight	GOVERN 3.1 MANAGE 1.3	§6.2, §9.1	Art. 14, Art. 16	DoD AI Eth. P.4
Secure Updates	MANAGE 2.4 MEASURE 3.2	§8.7, §10.2	Art. 12, Annex IV	STANAG 4586 §6.1
Auditability	MEASURE 2.5 GOVERN 4.2	§9.3, §10.3	Art. 17, Art. 20	DoD AI Eth. P.6
Robustness	MEASURE 2.2 MANAGE 2.2	§8.5, §9.2	Art. 9, Art. 15	IEEE 7010 §5.3

Derived from natural language source documents. Mappings are intended to be updated as standards evolve.

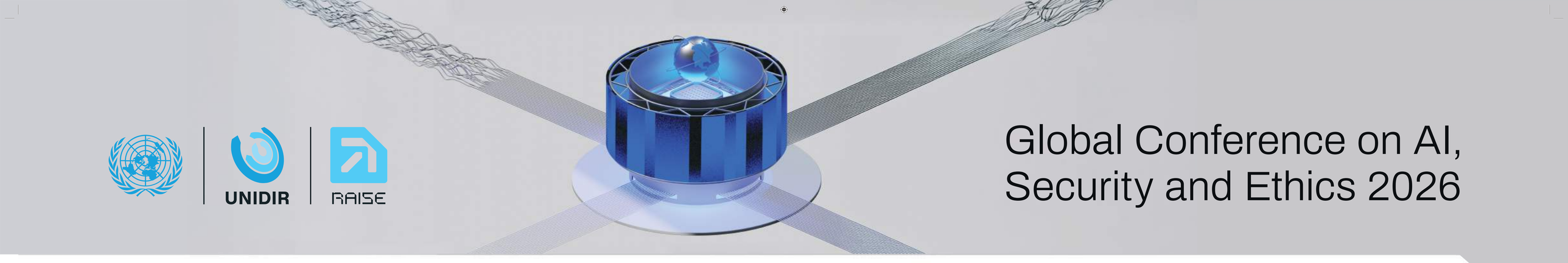
FUTURE DIRECTIONS & OPEN QUESTIONS

MULTILATERAL INTEROPERABILITY PILOT
Can allied nations exchange taxonomy-aligned assurance artifacts as a practical CBM? One possible research direction is exploring whether structured, machine-readable assurance reports could improve interoperability and comparability across institutions operating under different governance frameworks.

CONCEPT DRIFT DETECTION — AUTOMATIC RE-VERIFICATION
When a standard is revised, can the system automatically flag which existing assurance tags require re-verification? Machine-readable assurance representations could detect terminological shifts and alert compliance teams before gaps emerge.

EXTENSION BEYOND MILITARY AI
The same semantic structuring approach applies to civilian critical infrastructure — healthcare diagnostics, energy grid management, transportation control — where fragmented governance frameworks and inconsistent reporting structures create similar interoperability challenges.

OPEN COLLABORATION & DISCUSSION
This work is intended as an exploratory research proof of concept rather than a finalized operational platform. **Feedback, critique, collaboration ideas, and alternative approaches are all welcome, and we are fully open to discussion with researchers, policymakers, standards bodies, and technical practitioners interested in AI assurance interoperability.**



Global Conference on AI, Security and Ethics 2026

From Broad Ambition to Narrow Commitment: Bridging the Governance Gap in Military AI via Pipeline Auditing

Doelle Bhattacharya, Technical AI Safety & Governance Policy Researcher
Non-Trivial | doellebhattacharya@gmail.com

THE MULTILATERAL REGULATORY VACUUM

1. Introduction & Context

- Over three major REAIM summits, two UN General Assembly resolutions, and a rapidly expanding catalog of national AI strategies, the international community has failed to produce a single legally binding standard regulating the development or deployment of military artificial intelligence.
- The formal legal landscape remains fragmented: even the landmark European Union AI Act explicitly exempts military, defense, and national security applications from its scope.
- The Core Paradox: Global governance architectures have experienced an inversion—broadening significantly in scope and rhetorical ambition while narrowing sharply in actual political commitment and enforcement capability.

2. The Multilateral Polarization Trend

- The multilateral consensus is actively fracturing rather than consolidating. A direct assessment of recent state participation highlights a severe regression in diplomatic buy-in for non-binding responsible use frameworks:

Multilateral Milestone / Forum	State Endorsements / Status	Key Absentees / Regulatory Carve-Outs
REAIM Summit 2024 (Republic of Korea)	**61 States** signed the Blueprint for Action	Initial broad strategic alignment.
SIPRI Empirical Report (August 2025)	Evaluated systemic blind spots across operational systems	Documented compounding bias pathways.
REAIM Summit 2026 (Spain - Feb 2026)	**39 States** endorsed the framework (▼ 36% Decrease)	**United States, China, Russia, and Israel** (All absent).
EU AI Act (Enforced)	Regional Statutory Framework	**Explicitly Exempts** all military applications.

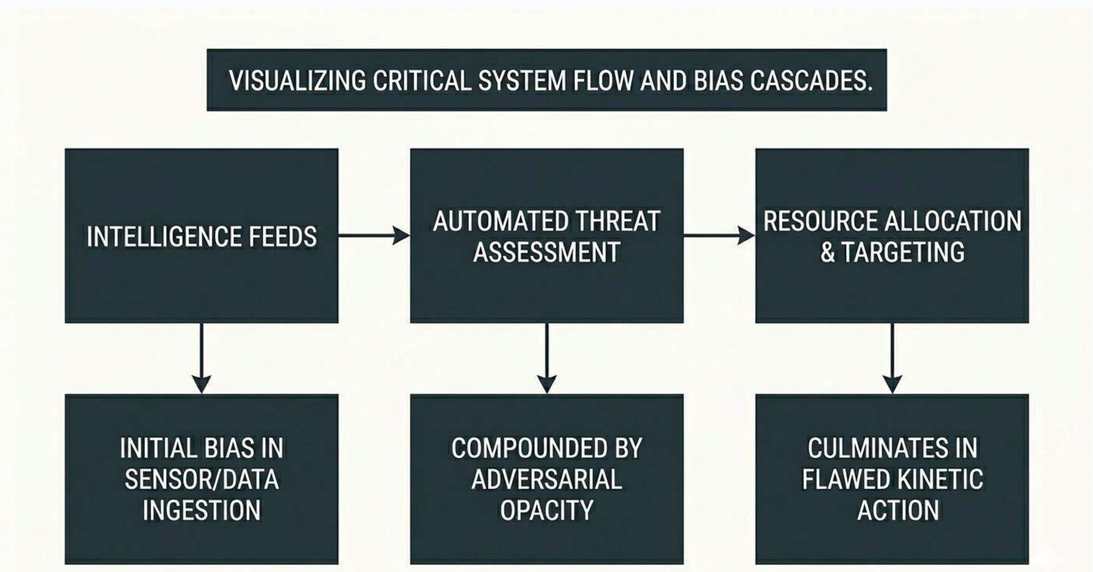
THE ANATOMY OF COMPOUNDING MILITARY BIAS

3. Expanding the Bias Taxonomy

- Traditional AI ethics frameworks limit algorithmic bias assessments to gender and race. However, empirical findings from the Stockholm International Peace Research Institute (SIPRI) released in August 2025 demonstrate that bias within operational military AI systems extends to highly complex, multi-dimensional categories:
  - **Disability & Age**: Algorithmic misclassifications of civilian movement profiles, gait styles, and processing speeds.
  - **Socioeconomic Class**: Structural skewing in automated regional analysis based on infrastructure density and visible wealth indicators.
  - **Cultural Context**: Dangerous failure modes when target-discrimination algorithms fail to parse local attire, regional customs, or non-Western behavioral patterns.

4. The Automation Bias Trap

- **Speed-Opacity Inversion**: The extreme execution speeds and deep optimization layers (opacity) of modern military Decision-Support Systems (DSS) induce severe automation bias.
- Human operators are left with an insufficient temporal window to evaluate, cross-verify, or overturn corrupted algorithmic recommendations.
- **The Pipeline Compounding Effect**: Rather than operating as isolated errors, undetected algorithmic biases act sequentially. Because military architectures are built on chained deployments, a baseline bias propagates aggressively across the operational pipeline
- **The Regulatory Deficit:** Despite this clear cascading threat vector, no existing multilateral or regional framework mandates rigorous, pipeline-level auditing for non-weapons military AI systems.



The Pipeline Compounding Effect (above)

THE PROPOSAL & TECHNICAL MECHANISMS

5. A Minimum-Viable Audit Standard (MVAS)

To bridge the gap between high-level ethics principles and concrete engineering, this poster outlines a verifiable, technical architecture for non-weapons military AI procurement. The standard introduces **three mandatory technical requirements**:

- **Requirement 1: Demographic Disaggregation**
  - *Mechanism*: Mandatory evaluation of error bounds and model outputs sliced across multi-dimensional protected characteristics (age, disability, cultural baseline data).
  - *Objective*: Prevents the masking of localized algorithmic failures behind generalized, aggregated accuracy metrics.
- **Requirement 2: Cross-System Propagation Testing**
  - *Mechanism*: Simulated stress-testing of downstream systems when fed intentionally corrupted or biased upstream data.
  - *Objective*: Quantifies exactly how bias compounds and self-amplifies across chained model deployments before field integration.
- **Requirement 3: Point-of-Procurement Disclosure**
  - *Mechanism*: Legally binding, standardized transparency reporting containing the exact results of demographic and propagation testing, delivered directly at the procurement phase.
  - *Objective*: Removes information asymmetries between defense contractors and military procurement officers.

6. Grounding: From Principle to Implementation

- This framework rejects the approach of outlining ethical standards purely as abstract moral virtues. Instead, it is designed from the functional perspective of a technical researcher working directly on the algorithmic mechanisms required to run these evaluations.
- By leveraging hands-on methodologies from **Reinforcement Learning from Human Feedback (RLHF) debiasing** and mathematical **algorithmic fairness frameworks**, we translate governance ideals into precise, machine-verifiable code pipelines.
- Bridging technical execution with international security policy represents a vital perspective that is missing from pure diplomatic tracks—offering defense authorities a concrete, actionable mechanism to enforce algorithmic accountability.

ABOUT THE AUTHOR

Doelle Bhattacharya is a student at Brookline High School and a Technical AI Safety & Governance Policy Researcher. She is second author on a peer-reviewed RLHF debiasing paper presented at NeurIPS and EMNLP 2025 workshops, and was a finalist for the Non-Trivial Research Foundations fellowship. As Managing Director of the Brookline Robotics Initiative, MIT Leadership Training Institute alum, and AI Research Fellow at Algoverse, she bridges hands-on algorithmic fairness engineering with multilateral security governance—a perspective largely absent from pure diplomatic tracks.

PUBLICATIONS

UNIDIR

Military Artificial Intelligence and Autonomy

SECURITY AND TECHNOLOGY

The Global Prism of Military AI Governance: Reflections from the 2025 Regional Consultations on Responsible AI in the Military Domain

SECURITY AND TECHNOLOGY

Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: An Evidence-Based Road Map for Future Policy Action

SECURITY AND TECHNOLOGY

Regional Perspectives on the Application of International Humanitarian Law to Lethal Autonomous Weapon Systems

SECURITY AND TECHNOLOGY

AI in the Military Domain: A briefing note for States

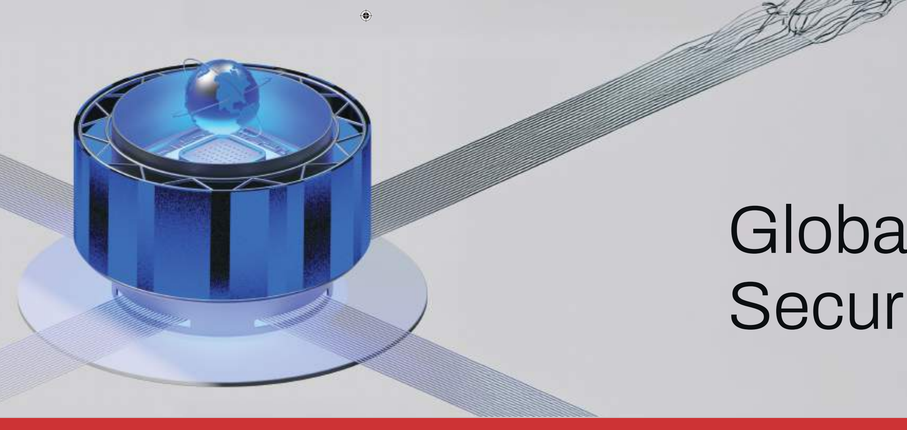
SECURITY AND TECHNOLOGY

The Interpretation and Application of International Humanitarian Law in Relation to Lethal Autonomous Weapon Systems

SECURITY AND TECHNOLOGY

Large Language Models and International Security: A Primer

UNIDIR.ORG/PUBLICATION



Global Conference on AI, Security and Ethics 2026

BUILDING TRUST IN AI FOR DEFENSE: *GOVERNANCE APPROACHES IN TÜRKIYE*

STRATEGIC CONTEXT

Türkiye is both a NATO member state with alliance governance obligations and a major developer/exporter of AI-enabled defense systems (Baykar TB2, KARGU, HAVELSAN AI). This dual position creates unique governance tensions between strategic autonomy and multilateral compliance.

SSB & PUBLIC-PRIVATE MODEL

Presidency of Defence Industries (SSB) coordinates Baykar, STM, HAVELSAN, ASELSAN, Roketsan. Effective for capability development; creates governance gaps where funder and customer are identical. No independent ethical review equivalent to UK DAIC or France's Comité d'Éthique.

NATO Principles Of Responsible Use Of AI In Defence (2021) | Endorsed By Türkiye As Alliance Member

1. **LAWFULNESS**
2. **RESPONSIBILITY & ACCOUNTABILITY**
3. **EXPLAINABILITY & TRACEABILITY**
4. **RELIABILITY**
5. **GOVERNABILITY**

IDENTIFIED GOVERNANCE GAPS

- No dedicated defence AI ethics/assurance body
- Oversight limited to legal compliance review
- No independent audit capacity separate from SSB
- Technology export governance for AI systems underdeveloped
- Limited inter-agency coordination mechanism

RECOMMENDATIONS

① Establish Defence AI Ethics & Assurance Board (DAEAB)
② Develop AI-specific technology export governance
③ Deepen multilateral CBM engagement (UN CCW GGE, NATO STO)
④ Create National Defence AI Coordination Council
⑤ Develop Defence AI Standardization Roadmap

Mariam Ahmed Mohamed Mansour maryahmedd1999@gmail.com
Kamal Tasiu kmlts256@gmail.com

18-19 June 2026
Palais des Nations,
Geneva, Switzerland

#AISE26

unidir.org