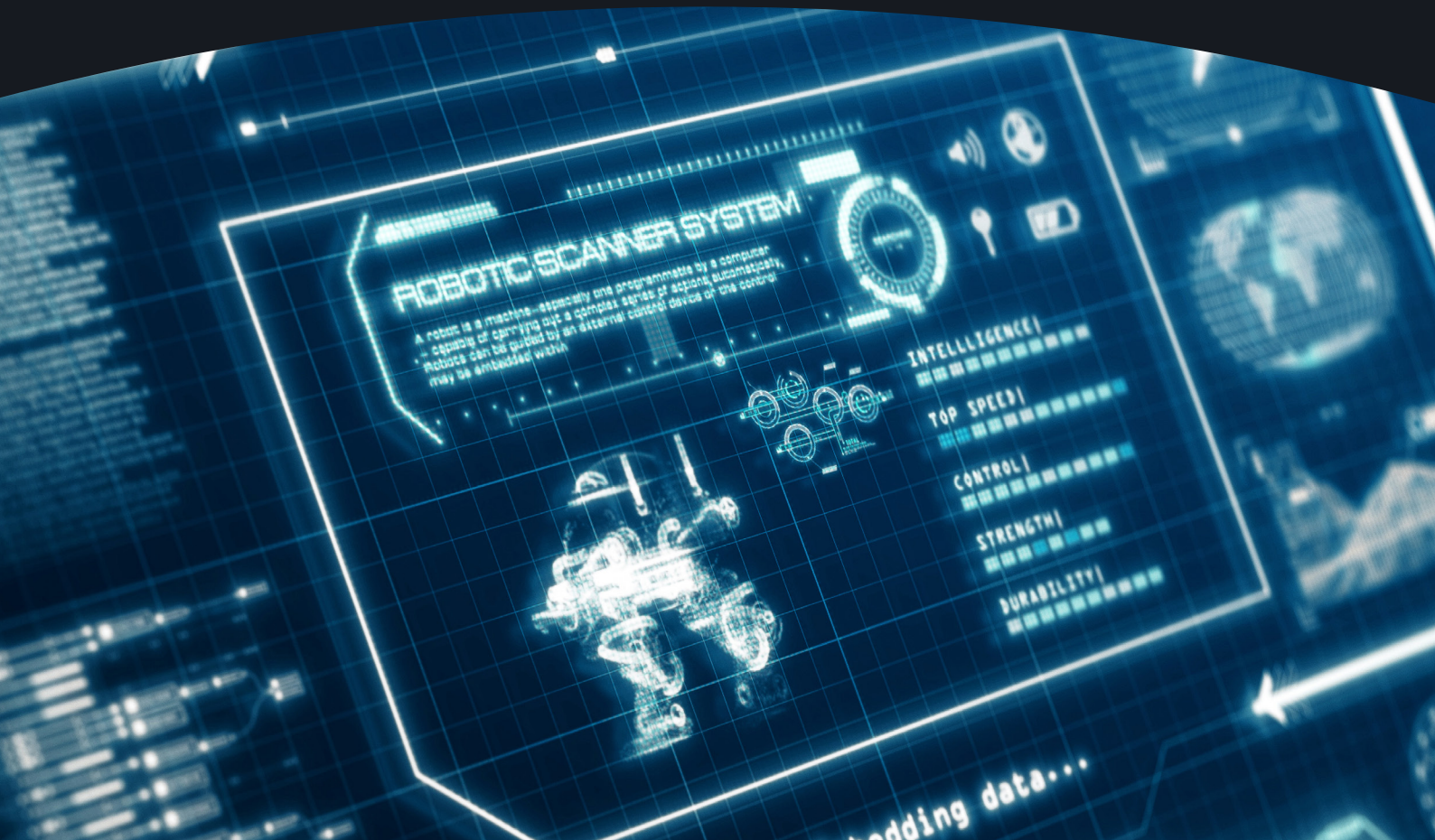
UNIDIR

From Cyberspace to Military Artificial Intelligence: Translating Norms of Responsible State Behaviour

BEYZA UNAL



Executive Summary

The report examines whether and how the voluntary, non-binding norms of responsible State behaviour developed in the field of information and communications technology (ICT) security can provide useful signposts for development of norms on artificial intelligence (AI) in the military domain. ICT norms should not be transferred directly to military AI, but lessons can instead be learned from the development and implementation of those norms. These can then be adapted to the distinct characteristics of AI. The important limits to the comparison between the two fields mean that any future normative considerations on AI in the military domain would need to be tailored to the technology's distinctive characteristics, operational contexts, and implications for international peace and security.

There are three areas of convergence between ICT security and AI in the military domain: the relationship between responsibility-based and capability-based policy approaches; the central role of the private sector; and the dual-use nature of the underlying technologies. Depending on the capability or application concerned, voluntary norms and legally binding measures may be developed successively, operate concurrently or reinforce one another. In both approaches, there is also value in transparency and confidence-building measures (TCBMs), risk-reduction measures, and verification-related practices.

Using the 11 voluntary norms of responsible State behaviour in ICT security as analytical reference points, possible areas for future normative consideration on AI in the military domain can be identified. These include international cooperation and capacity-building; assessment of AI-enabled incidents; prevention of misuse by State and non-State actors; respect for international law and human rights; protection of civilian infrastructure; assistance and incident response; supply chain integrity; and responsible vulnerability disclosure.

Convergence is emerging among Member States around the applicability of international law throughout the life cycle of AI in the military domain, the maintenance of control and human judgment in the use of force, human responsibility and accountability, protection of civilians in armed conflict, and the value of international cooperation. At the same time, significant differences remain regarding shared understanding around terminology, the desirability of voluntary or binding norms, the identification of red lines, and the appropriate institutional forum and modalities.

High-risk applications of AI, such as lethal autonomous weapons systems and the potential integration of AI into nuclear weapon, are addressed in distinct multilateral processes but raise common concerns about control, human judgment, miscalculation, and escalation.

Experience from ICT security demonstrates the value of sustained, inclusive and incremental norm-building. However, the pace of AI development requires international discussions to advance more rapidly. Further work is needed on terminology, behavioural intentions, structural engagement between Member States and industry, technical assurance, TCBMs, and capacity-building.

Acknowledgements

Support from UNIDIR's funders provides the foundation for all of the Institute's activities. This study was produced by UNIDIR's Security and Technology Programme (SECTEC), which is supported by the Governments of Germany, Italy, the Netherlands, Switzerland, and Microsoft.

The author would like to warmly thank United Nations Institute for Disarmament Research (UNIDIR), in particular Dr. Giacomo Persi Paoli, for his support throughout the publication of this paper. The author is also grateful to Dr. Andraz Kastelic (UNIDIR), Kathleen Lawand, Li Zhou (OHCHR), Michael Karimian (Microsoft), Dr. Vincent Boulanin (SIPRI) and Dr. Yasmin Afina (UNIDIR) for their review and constructive inputs. Finally, the author would like to thank her daughter, Mina, whose patience as a one-year-old, made it possible for her mom to write this paper.

About UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institution within the United Nations. One of the few policy institutes worldwide focusing on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, UNIDIR assists the international community in developing the practical, innovative ideas needed to find solutions to critical security problems. For more information, visit unidir.org.

Notes

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

Citation

Beyza Unal, *From Cyberspace to AI: Translating Norms of Responsible State Behaviour in Cyberspace into AI in the Military Domain*, Geneva: UNIDIR, 2026. https://doi.org./10.37559/SECTEC/26/CR/07.

Author

Dr. Beyza Unal

Head of the Science, Technology and International Security Unit of the United Nations Office for Disarmament Affairs (UNODA)

Beyza Unal specializes in the interaction between emerging technology applications and their impact on international peace and security. Before joining UNODA, Dr. Unal was Deputy Director of the International Security Programme at Chatham House. She formerly worked in the Strategic Analysis Branch at the Allied Command Transformation (ACT) of the North Atlantic Treaty Organization (NATO), taught international relations at Old Dominion University, and served as an international election observer during the 2010 Iraqi parliamentary election. Dr. Unal has a doctoral degree from Old Dominion University, Virginia, United States. She has published numerous articles on international security, including on nuclear weapon policy, cybersecurity, outer space security, critical infrastructure protection and artificial intelligence.

Acronyms and Abbreviations

AI	Artificial intelligence
AISIRT	AI Security Incident Response Team
CBM	Confidence-building measure
CCW	(Convention on) Certain Conventional Weapons
CERT	Computer emergency response team
COP	Community of practice
CSIRT	Computer security incident-response team
GGE	Group of Governmental Experts
ICT	Information and communications technology
ICRC	International Committee of the Red Cross
IHL	International humanitarian law
IHRL	International human rights law
LAWS	Lethal autonomous weapon systems
LLM	Large language model
POC	Point of contact
OECD	Organisation for Economic Co-operation and Development
OHCHR	Office of the United Nations High Commissioner for Human Rights
OT	Operational technologies
REAIM	Responsible Artificial Intelligence in the Military Domain
SBOM	Software bill of materials
SOP	Standard operating procedure
TCBM	Transparency and confidence-building measure
POC	Points of contact
TEVV	Testing, evaluation, verification and validation
TTP	Tactics, techniques and procedures

Contents

Credit: © DKosig/iStock

1. Introduction

In 2015, a United Nations group of governmental experts (GGE) articulated 11 voluntary and non-binding norms of responsible State behaviour in the security and use of information and communications technologies (ICTs).[1] These voluntary norms aim to reduce risks to international peace and security, increase trust among States and contribute to the prevention of conflict in the ICT domain.

In contrast, in the case of artificial intelligence (AI) in the military domain, there are no multilaterally agreed norms (legally binding or non-binding) despite growing momentum and increased engagement by both States and other stakeholders. State views on this question appear to fall along a spectrum. Some consider the development of voluntary norms unnecessary, preferring a legally binding instrument instead. Others do not reject voluntary norms in principle but emphasize the need to first build shared understanding and common approaches before moving towards development of norms. A third group of States has called for the development of norms on AI in the military domain, stressing that such efforts are needed before technology advances in ways that may outpace governance.

---

1 Throughout this report, the term "norms" is used in three distinct ways. When referring to the agreed language of the United Nations intergovernmental processes, the original terminology is retained with the use of the expression "voluntary and non-binding norms". When referring to norms as a broader governance concept, irrespective of whether they are legally binding or non-binding, the term "norms" is used without qualification. Finally, when discussing non-legally binding normative measures more generally (rather than referring specifically to the agreed ICT security language), the term "voluntary norms" is used.

Without any prejudice into the first two perspectives, this report contributes to the discussion on the potential development of voluntary and non-binding norms by examining what such norms could look like should States decide to develop them in the future.

Standalone multilateral discussions on AI in the military domain are currently taking place within the First Committee of the United Nations General Assembly.[2] To facilitate these discussions, States and stakeholders submitted their views to a report of the Secretary-General on "Artificial intelligence in the military domain and its implications for international peace and security" in 2025. In June 2026, States and stakeholders also engaged in a three-day informal exchange on AI in the military domain, sharing their views on the Secretary-General's report.[3]

The Secretary-General's report captures opportunities and challenges, provides a catalogue of existing and emerging normative proposals, and surveys existing initiatives in the field of AI in the military domain.[4] The report concludes with the Secretary-General's observations and conclusions, including a recommendation for States to take concrete steps to establish a dedicated and inclusive process to address AI in the military domain and its implications for international peace and security.

In parallel, research institutions and stakeholder communities have begun to explore concrete pathways forward. For example, a recent UNIDIR publication provides recommendations on the potential next steps, including developing a set of core principles of responsible AI in the military domain, confidence-building measures (CBMs) for military applications of AI and capacity-building programmes.[5]

Against this backdrop and amid widespread concerns that developments in the field of AI are outpacing existing governance frameworks, this report examines whether existing voluntary norms in ICT security can serve as a useful signpost for norm development in the military AI domain.[6] The report acknowledges at the outset that norms to be developed in the field of military AI need to be tailored to its unique characteristics. Moreover, there are limits to the analogy: lessons in ICT security may not translate cleanly to AI in the military domain. For instance, unlike in ICT security – where voluntary norms were developed after decades of technological maturation and operational experience – military AI remains an evolving field, with capabilities and concepts still taking shape. Despite this limitation, it is argued here that important lessons can be drawn from State practice and experience in responsible behaviour in ICT security.

---

2 The General Assembly adopted its first resolution on AI in the military domain on 31 December 2024. General Assembly, A/RES/79/239, 2024, https://docs.un.org/A/RES/79/239.

3 On the informal exchanges, see "Informal exchanges on artificial intelligence in the military domain", UNODA Meeting Place, 2026, https://meetings.unoda.org/informal-exchanges-on-artificial-intelligence-in-the-military-domain-2026.

4 General Assembly, "Artificial intelligence in the military domain and its implications for international peace and security", Report of the Secretary-General, A/80/78, 5 June 2025, https://docs.un.org/A/80/78.

5 UNIDIR Security and Technology Programme, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: An Evidence-Based Road Map for Future Policy Action* (Geneva: UNIDIR, 2025), https://unidir.org/publication/artificial-intelligence-in-the-military-domain-and-its-implications-for-international-peace-and-security-an-evidence-based-road-map-for-future-policy-action.

6 This report uses the word "signpost" rather than "model" or "template" because norms in ICT security cannot be applied directly to AI in the military domain. Instead, existing norms in ICT security could be considered as a marker or indicator for the development of norms in AI in the military domain.

While there is no single universal definition of artificial intelligence, it relates to machines with the ability to learn, solve problems, make predictions, take decisions and perform tasks that are considered to require human intelligence. This report uses the term "artificial intelligence" when referring also to the subset of AI techniques that includes, for example, machine learning, deep learning, generative AI and agentic AI.

The term "military domain" can also be interpreted differently.[7] For some States, the military domain may include a broad range of security- and law enforcement-related functions, such as combating organized crime or policing. Others interpret it more narrowly as activities undertaken by armed forces in the context of armed conflict, including pre-conflict, conflict and post-conflict situations.[8] In this report, the military domain refers to both operations and activities in armed conflict and also the use of force in security and law enforcement contexts (including terrorism, organized crime, border security etc.) and other activities that may require military assistance, such as in the context of disaster relief or rescue operations. Overall, this report also acknowledges the need for terminological clarification in the multilateral forums on what constitutes the military domain; an item that was featured in the programme of the 2026 informal exchanges on AI in the military domain.

This report is primarily written for an expert audience, including the diplomatic community, policymakers, academia, industry and other technical stakeholders. It first provides (in Section 2) background on the 11 voluntary norms of responsible State behaviour in ICT security. It then highlights (in Section 3) convergences between ICT technologies and AI in the military domain through the lens of voluntary norm development. The report goes on to use (in Section 4) existing voluntary norms in ICT security as a signpost for AI in the military domain. It concludes (in Section 5) with final observations including research suggestions. In its methodology and a primary resource, the report relies on the views of States and stakeholders submitted in the Secretary-General's report.[9] The report also provides (in Appendix I) principles identified in existing State-led initiatives in order to have a comprehensive approach to norm development.

---

7 For a detailed analysis of the notion of "the military domain", see UNIDIR Security and Technology Programme, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security*.

8 UNIDIR Security and Technology Programme, *Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security*.

9 The submissions included 31States views submitted on time, as well as an additional 10 submissions received after the deadline and made available on the Secretariat's website.

Credit: gorodenkoff/iStock

2. Background on voluntary norms of responsible State behaviour in ICT security

Information security has been on the United Nations agenda since the Russian Federation first introduced a draft resolution in 1998 in the First Committee of the General Assembly. Since then, States have engaged in intergovernmental discussions to assess information and communications technology related threats and identify cooperative measures to address them.[10]

Intergovernmental processes on ICT security have taken place primarily in two formats: groups of governmental experts (GGEs) and open-ended working groups (OEWGs).[11] Since 2004, six GGEs have examined existing and potential threats from the use of ICTs and possible cooperative measures to address them. Two OEWGs were also convened to broaden participation and advance discussions on developments in the field of ICTs, with the second OEWG concluding its last session in July 2025.[12]

---

10 The topics discussed under ICT threats include, among others, ransomware, integrity of the supply chain, commercially available ICT intrusion capabilities.

11 The coexistence of these two formats reflects differing State preferences regarding the size, inclusiveness and negotiating dynamics as well as evolving views on how best to advance consensus on ICT security.

12 Samuele Dominioni and Giacomo Persi Paoli (eds.), *Unpacking the United Nations Open-Ended Working Group on ICT in the Context of International Security (2021–2025)* (Geneva: UNIDIR, 2026), https://unidir.org/publication/unpacking-the-united-nations-open-ended-working-group-on-ict-in-the-context-of-international-security-2021-2025.

In 2026, the Global Mechanism on Developments in the Field of ICTs in the Context of International Security and Advancing Responsible State Behaviour in the Use of ICTs (Global Mechanism on ICT Security) commenced its work.

A central outcome of the GGE- and OEWG-led processes was the development in 2015 of 11 voluntary, non-binding norms of responsible State behaviour in ICT security by the fourth GGE (see Table 2.1).[13] These voluntary, non-binding norms were welcomed by consensus through a resolution adopted by the United Nations General Assembly in December 2015, which calls upon States to be guided in their use of ICTs by the 2015 GGE report.[14]

Table 2.1. Existing norms of responsible State behaviour in ICT security

	NORM	EXPLANATION
a	International cooperation	Consistent with the purposes of the United Nations, including to maintain international peace and security, States should cooperate in developing and applying measures to increase stability and security in the use of ICTs and to prevent ICT practices that are acknowledged to be harmful or that may pose threats to international peace and security.
b	Consider all relevant information	In case of ICT incidents, States should consider all relevant information, including the larger context of the event, the challenges of attribution in the ICT environment and the nature and extent of the consequences.
c	Prevent misuse of ICTs in your territory	States should not knowingly allow their territory to be used for internationally wrongful acts using ICTs.
d	Exchange information to prevent crime and terrorism	States should consider how best to cooperate to exchange information, assist each other, prosecute terrorist and criminal use of ICTs and implement other cooperative measures to address such threats. States may need to consider whether new measures need to be developed in this respect.
e	Respect human rights and privacy	States, in ensuring the secure use of ICTs, should respect Human Rights Council resolutions 20/8 and 26/13 on the promotion, protection and enjoyment of human rights on the Internet, as well as General Assembly resolutions 68/167 and 69/166 on the right to privacy in the digital age, to guarantee full respect for human rights, including the right to freedom of expression.

13 General Assembly, Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, A/70/174, 22 July 2015, https://docs.un.org/A/70/174.

14 General Assembly, A/RES/70/237, 2015, https://docs.un.org/A/RES/70/237.

Table 2.1. continued

	NORM	EXPLANATION
f	Do not damage critical infrastructure	A State should not conduct or knowingly support ICT activity contrary to its obligations under international law that intentionally damages critical infrastructure or otherwise impairs the use and operation of critical infrastructure to provide services to the public.
g	Protect critical infrastructure	States should take appropriate measures to protect their critical infrastructure from ICT threats, taking into account General Assembly resolution 58/199 on the creation of a global culture of cybersecurity and the protection of critical information infrastructures.
h	Respond to requests for assistance	States should respond to appropriate requests for assistance by another State whose critical infrastructure is subject to malicious ICT acts. States should also respond to appropriate requests to mitigate malicious ICT activity aimed at the critical infrastructure of another State emanating from their territory, taking into account due regard for sovereignty.
i	Ensure supply chain security	States should take reasonable steps to ensure the integrity of the supply chain so that end users can have confidence in the security of ICT products. States should seek to prevent the proliferation of malicious ICT tools and techniques and the use of harmful hidden functions.
j	Report ICT vulnerabilities	States should encourage responsible reporting of ICT vulnerabilities and share associated information on available remedies to such vulnerabilities to limit and possibly eliminate potential threats to ICTs and ICT-dependent infrastructure.
k	Do not harm emergency response teams	States should not conduct or knowingly support activity to harm the information systems of the authorized emergency response teams (sometimes known as computer emergency response teams or cybersecurity incident response teams) of another State. A State should not use authorized emergency response teams to engage in malicious international activity.

In addition to the development of voluntary, non-binding norms, States have given attention to their implementation and to the potential development of new norms. The 2013 GGE report notes that, "Given the unique attributes of ICTs, additional norms could be developed over time".[15] Also addressing this issue, the 2021–2025 OEWG reiterated that new norms could be developed in the future and that "the further development of norms, and the implementation of existing norms were not mutually exclusive but could take place in parallel".[16] States also gave particular attention to the implementation and potential development of new norms. The implementation of norms was discussed by the 2019–2021 OEWG.

Rather than creating new obligations, the voluntary, non-binding norms of responsible State behaviour in ICT security are grounded in "existing international norms and commitments".[17] In this regard, the 2021 GGE report indicates that the norms and international law are complementary to each other and that the norms "do not seek to limit or prohibit actions that is otherwise consistent with international law".[18]

The voluntary norms on ICT security are composed of both positive commitments that recommend States to take particular action and negative commitments that recommend States not to engage in certain behaviours. Of the 11 norms, 8 articulate positive commitments (a, b, d, e, g–j), while 3 contain negative commitments (norms c, f and k).[19] This balance between positive and negative commitments is a useful analytical lens for assessing what types of commitment may be feasible in the military AI context.

Taken together, the voluntary norms on ICT security illustrate how States have addressed international cooperation, CBMs, critical infrastructure protection and attribution-related challenges in a complex and rapidly evolving domain. The early recognition of the security risks associated with ICTs provides an important precedent for current discussions on military applications of AI, illustrating how States conceptualized emerging technology challenges. The evolution from expert-driven to more inclusive formats is also analytically relevant for AI in the military domain. Moreover, the ICT experience demonstrates that, even in a challenging international security environment, States can reach consensus on a non-binding normative framework when there is sufficient political will. This lesson is directly relevant to ongoing discussions on AI in the military domain, despite divergent views on how discussions should proceed. Lastly, the dual-track approach – simultaneously implementing existing voluntary norms while considering future development of new ones – offers a potential model for addressing fast-evolving technologies such as AI in the military domain.

---

15 General Assembly, Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, A/68/98, 24 June 2013, https://docs.un.org/A/68/98, paragraph 16.

16 General Assembly, Final report of the open-ended working group on security of and in the use of information and communications technologies 2021–2025, A/80/257, 24 July 2025, https://docs.un.org/A/80/257, paragraph 36(d).

17 General Assembly, A/70/174, paragraph 11.

18 General Assembly, Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, A/76/135, 14 July 2021, https://docs.un.org/A/76/135, paragraph 15.

19 The consideration of positive and negative commitments goes beyond ICT security. It is often considered in the development of international law and arms control treaties.

Credit: © koto_feja/iStock

3. Convergences of ICT security and AI in the military domain

3.1 Responsibility- and capability-based approaches

In recent years, discussions among States around emerging technologies, including on information and communications technologies and artificial intelligence, have taken place under two partly overlapping approaches to technology governance: responsibility-based approaches (also known as behavioural approaches) and capability-based approaches. These two approaches often represent a spectrum, rather than mutually exclusive categories (see Table 3.1).

Responsibility-based approaches often focus on *how* technologies are used and *who* is accountable for their use. In contrast, capability-based approaches emphasize which technologies or *what* types of capability are developed, possessed or deployed and *which* should be prohibited or regulated. In practice, however, many governance measures combine elements of both: responsibility-based commitments may include restrictions on particular types of use, while capability-based arrangements may also contain transparency, risk-reduction and confidence-building provisions.

Responsibility-based approaches focus on regulating *behaviour* and *use* of technology, rather than technology itself. States, in this regard, are trusted to exercise self-restraint, often through the development and implementation of voluntary and non-binding norms, rules and principles. Other behavioural approaches include developing political commitments, guidelines, or codes of conduct or practice. States often move to convergence on the topic incrementally through voluntary measures. However, behavioural expectations can also be reflected in legally binding instruments,

including where treaties codify duties of restraint, due diligence, notification measures, transparency and compliance with international humanitarian law (IHL).

Capability-based approaches, in contrast, regulate the development, possession or deployment of specific capabilities through legally binding measures, prohibitions or limitations (e.g., restrictions on the nature of weapon payload, size, weight, range of platform). Traditional arms control regimes often follow this model. Prominent examples include the prohibition of the proliferation of nuclear weapons or the bans on the use and possession of biological and chemical weapons.

While responsibility-based approaches are often technology agnostic and focus on what specific actors (e.g., States, industry, operators etc.) should or should not do, capability-based approaches are technology centric and provide clear parameters to the actors around what technologies and capabilities they can and cannot design, develop, deploy or use. Yet even this distinction can blur. For example, a norm that requires control and human judgment over the use of force may be framed as a responsibility-based commitment, but it can also have capability-related implications if certain systems are considered unacceptable because they cannot be operated under such control. Similarly, a prohibition or limitation on capability may still require behavioural commitments (e.g., voluntary reporting) in order to be implemented effectively.

There are other elements in the disarmament toolbox that either sit in between responsibility-based and capability-based approaches or can serve both approaches. Examples include transparency and confidence-building measures (TCBMs), verification, and risk reduction. TCBMs are often considered as non-legally binding in nature, integrated into arms control arrangements; however, States could equally incorporate TCBMs as part of a legally binding arms control agreement. Historical examples demonstrate that legally binding agreements may incorporate transparency, notification or information-sharing measures that perform TCBM functions (e.g., the Treaty on Conventional Armed Forces in Europe).

Risk reduction can also be a measure of both voluntary norms and legally binding measures. Non-legally binding political commitments may perform risk-reduction functions that later support, inform or run alongside legally binding arrangements.

Meanwhile, verification mechanisms are most commonly associated with legally binding instruments, although their effectiveness often depends on preparatory or complementary non-legally binding measures. Voluntary exchanges of information, reporting practices, technical consultations or data-sharing practices can help establish baselines, build technical familiarity, and develop conditions that make formal verification more feasible or credible.

In practice, States rarely rely exclusively on one approach; but their national policy positions often tilt between these two approaches and at times the two approaches can overlap to some extent.

Both responsibility- and capability-based approaches could complement each other through a third, hybrid approach. For instance, non-binding norms could be developed to pave the way for areas where legally binding obligations could subsequently be sought; or non-binding norms could themselves become binding through customary international law. States could also rely on both non-binding norms and legally binding measures at the same time, depending on the application of technology. Lastly, States can develop norms of behaviour by elaborating international law.

Table 3.1. Indicative distinctions and overlaps between responsibility-based and capability-based approaches

	RESPONSIBILITY-BASED APPROACH	CAPABILITY-BASED APPROACH	AREAS OF OVERLAP/ HYBRID PRACTICE
Framing	Behaviour of or use by an actor	Development, possession, transfer, deployment or use of a technology or capability	Some behavioural commitments may have capability-based effects; some capability limits may require behavioural implementation measures
Primary focus	How actors should or should not behave	What capabilities should be permitted, limited, prohibited or controlled	Areas where responsible use commitments co-exist with limitations or restrictions on specific high-risk applications
Legal form	Often expressed through non-binding norms, rules, principles, political commitments, declarations, codes of conduct or guidelines	Often expressed through legally binding instruments, prohibitions, limitations, export controls or compliance measures	TCBMs Verification measures Risk reduction
Institutional settings at the United Nations	Often discussed through the General Assembly, often in GGEs or OEWGs	Often discussed through the Convention on Certain Conventional Weapons (CCW), the Conference on Disarmament, the General Assembly	Institutional forum is not determinative: the General Assembly can consider both responsibility- and capability-based approaches, and treaty bodies can include behavioural approaches
Advantages	Flexible and adaptable to technological change; can support incremental convergence	Can provide clearer limits; may strengthen compliance expectations; enforceable; adaptable to changes through reviews, such as review conferences	Hybrid approaches can combine flexibility with clearer commitments, including through review mechanisms, reporting or TCBMs
Disadvantages	May have weak compliance assurance	May face definitional, verification and adaptation challenges, especially for dual-use technologies	May still face political disagreement over sequencing or implementation

The relationship between these approaches may therefore be successive, concurrent or mutually reinforcing.[20] In some contexts, non-legally binding measures may build shared understanding before a treaty is politically or technically feasible.

In the field of ICT security, while responsibility-based approaches have been at the forefront of United Nations discussions, elaboration of new legally binding obligations was also discussed in the 2021–2025 OEWG on ICT security and the July 2026 plenary session of the Global Mechanism on ICT Security.[21] In particular, a group of like-minded States, led by the Russian Federation, has advocated "to conclude a legally binding multilateral treaty within the United Nations" to prevent conflict in the ICT domain and promote peaceful uses of ICTs.[22] With the establishment of the Global Mechanism on ICT Security in December 2025, the substantive focus will continue to encompass not only the implementation of existing voluntary norms but also discussions on the development of new voluntary norms, as well as discussion on legally binding obligations.

The coexistence of normative and legal tracks evolving in parallel reflects a broader pattern in international security governance. While it is beyond the scope of this report, similar dynamics – where voluntary norms or guiding principles and legally binding measures are pursued either concurrently or back-to-back – have also been observed in discussions on outer space security and lethal autonomous weapon systems (LAWS).[23]

A similar dynamic is evident in the field of AI in the military domain. While some States are calling for responsible use of AI in the military domain, several States oppose the notion of "responsibility" due to its subjective interpretation and on the grounds that "responsible application does not constitute a universally recognized category of international law".[24] Nevertheless, the progress achieved in ICT security, especially with the establishment of the Global Mechanism – despite divergences of views among States on the term "responsibility" – suggests that responsibility-based frameworks can generate convergence over time. It may also suggest that, through non-binding norms, States can still converge around concrete commitments, such as refraining from certain conduct, protecting critical infrastructure, maintaining control and human judgment or ensuring accountability.

---

20 The author thanks Katherine Prizeman for providing this categorization during an internal discussion at UNODA.

21 See "Global Mechanism on ICTs in the Context of International Security – Plenary", UNODA Meeting Place, 2026, https://meetings.unoda.org/meeting/76825/documents.

22 Russian Federation, "Updated concept of the Convention of the UN on Ensuring International Information Security", Working Paper, 2024, https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-_(2021)/ENG_Concept_of_convention_on__ensuring_international_information_security.pdf.

23 In the area of LAWS, States agreed on 11 non-binding guiding principles in 2019. Current discussion involves making meaningful progress to move towards negotiations on a new legally binding instrument to set clear prohibitions and regulations on autonomous weapon systems.

24 Russian Federation, "Position of the Russian Federation on the prospects of addressing the issue of military use of artificial intelligence in the military domain at the international level, including within the United Nations framework", Working Paper (unofficial translation), 12 June 2026, https://unodaweb-meetings.unoda.org/public/2026-06/[ENG] WP of the Russian Federation.pdf. For example, the Islamic Republic of Iran objects to the terminology "responsible application" as being too abstract. See General Assembly, A/80/78, 42.

Taken together, these examples suggest that the nature of specific AI capabilities and their use cases are likely to shape whether States will favour voluntary norms, legally binding measures or a combination of both in the future. For instance, while there is no universal agreement, some high-risk use cases whereby systems can identify, select and engage targets without control and human judgment are widely considered by the international community to be off-limits.[25] Moreover, the ongoing debate on human control provides valuable insights into other domains grappling with the same question, including the potential integration of AI into nuclear command, control and communications (NC3) systems.[26]

3.2 The role of the private sector

A second area of convergence between ICT security and military AI applications lies in the central role of the private sector. In both fields, States do not exercise exclusive control over the design and development of the relevant technologies. This makes effective norm implementation in areas related to safety, security and reliability of AI systems dependent not only on States but also on non-State actors, namely the private sector.[27]

In the realm of ICT security, much of the digital infrastructure and related services are developed, owned, operated and used by private entities, giving cyberspace a uniquely multi-stakeholder character. As a result, effective implementation of voluntary norms has depended on sustained public–private cooperation, rather than purely State-centric regulation. The private sector provides technical expertise, and it contributes to best practices in threat detection, incident response and in building resilience in the digital infrastructure.

The lessons from ICT security suggest that norms for AI in the military domain cannot be developed or implemented in isolation from the private sector. Lessons such as the importance of public–private partnerships, coordinated incident response, the role of computer emergency response teams (CERTs) across sectors, and multi-stakeholder cooperation provide transferable insights from ICT security to AI in the military domain.

Engagement with industry should not be understood as a substitute for State responsibility. Rather, it should be seen as a means of improving technical quality, feasibility and implementation of State-led discussions.

---

25 While States have not agreed on a single definition of control, the 2026 GGE on LAWS agreed by consensus that, to uphold compliance with international humanitarian law applicable in armed conflict, and in particular the principles and requirements of distinction, proportionality and precautions in attack, "control and human judgment with regard to LAWS are needed". The report further recognizes that human beings may exercise such control directly or indirectly, including through measures taken before and/or during the use of LAWS, and that humans need to exercise judgment to determine and implement the measures necessary to ensure compliance with IHL. See CCW Convention, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapon Systems, Advance Copy, Report of the 2024-2025-2026 sessions, 14 September 2026, https://meetings.unoda.org/meeting/79329/documents?f%5B0%5D=document_type_documents%3AFinal%20reports.

26 General Assembly, A/RES/80/23, 2025, https://docs.un.org/en/A/RES/80/23.

27 The role of the private sector in designing and developing AI applications safely and securely does not shift the responsibility from States to ensure that technologies designed or developed under their jurisdiction meet applicable safety and security standards.

The private sector plays a critical role in reducing risks to international peace and security through responsible design and development of AI systems.[28] The United Nations Guiding Principles on Business and Human Rights is a good starting point to incorporate peace and security implications of AI. The Guiding Principles clearly articulate the State duty to protect, the corporate responsibility to respect human rights, and the individual's or group's right to access remedy. Moreover, while the Global Digital Compact – adopted by the General Assembly in 2024 alongside the Pact for the Future – is addressed to the civilian domain, it calls for digital technology companies and developers to respect international human rights law and principles, including through human rights due diligence and impact assessments throughout the technology life cycle.[29] The Office of the United Nations High Commissioner for Human Rights (OHCHR) has also issued guidance to businesses to identify and assess human rights-related risks.

There may be areas where State commitments can be implemented through industry practice and market incentives, and others where industry approaches can complement State practices that are implemented through procurement authorities or certified third-party organizations.

The private sector also bears a responsibility to establish safeguards that reduce the risk of AI systems being used outside the parameters of intended use or contractual user agreements. This practice also helps with responsible sales and procurement of these models. Industry-led practices, such as the use of evaluation metrics (often referred as "evals"), safety and security testing, expert-led red teaming, adversarial testing (often referred as "jailbreaking"), and ongoing research on AI safety and security, demonstrate how technical safeguards can complement norms. Many AI companies stress test a system or model using adversarial inputs to evaluate its robustness and resilience.[30] They also develop methods to improve the transparency and interpretability of AI systems, allowing human operators to trace and review AI-enabled decisions. These efforts aim to move away from "black box" systems (where the inner workings of the system are hidden or too complex to interpret) to a transparent one, enabling human oversight – an issue of particular relevance for military AI applications.

These practices illustrate how private sector-led mechanisms can support the operationalization of norms related to safety, reliability and responsible deployment of AI systems. However, industry engagement in AI in the military domain remains limited in multilateral discussions. This creates a gap between technical practice and policy deliberation. Greater structural engagement with industry could therefore help States better understand the technical–policy interface, assess the feasibility of proposed

---

28 A joint initiative between UNIDIR and OHCHR examines how existing international law, norms and principles, including the Guiding Principles, could be translated into responsible industry behaviour in the context of AI in the military domain. For more information, see UNIDIR, "Framework for responsible industry behaviour for AI in the military domain", https://unidir.org/framework-of-responsible-industry-behaviour-for-ai-in-the-military-domain.

29 Global Digital Compact, annexed to General Assembly, A/RES/79/1, 2024, https://docs.un.org/A/RES/79/1, annex I, paragraphs 22–25.

30 See for instance Anthropic's approach to "frontier threats red teaming", whereby Anthropic works with experts to identify biology related risks to national security. Anthropic, "Frontier threats red teaming for AI safety", 26 July 2023, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

principles or political commitments and develop policy solutions that are informed by current industry practices and limitations. Such engagement could also benefit industry. It could provide companies with a greater understanding of multilateral discussions, which in turn could help companies design products, safeguards and governance processes in ways that are more attentive to international peace and security concerns.

Concrete examples that the AI private sector could adopt and that States can reinforce through procurement and partnerships include:

- Testing, evaluation, verification and validation (TEVV) expectations from industry, red teaming, and continuous evaluation for military-relevant use cases
- Documentation and traceability across models, data and updates
- Responsible vulnerability-disclosure pathways adapted for sensitive contexts
- Criteria of reporting serious incidents and secure decommissioning practices to reduce diversion and misuse
- Responsible sales and end-use controls where feasible, aligned with due diligence under the United Nations Guiding Principles on Business and Human Rights

3.3 Dual-use technology

A third convergence between ICT security and AI lies in their inherently dual-use nature. Technologies originally developed for civilian or commercial purposes can be repurposed or exploited by States or non-State actors for harmful purposes or military applications.

The dual-use challenge has long been recognized in United Nations discussions. For example, during the April 2025 session of the Disarmament Commission's Working Group II, many States emphasized that emerging technologies are inherently neutral, while their applications can either advance peaceful objectives or pose significant risks to international peace and security.[31] This framing aligns closely with ICTs and AI in the military domain.

In the field of ICT security, existing norms – such as the expectation that States prevent the misuse of ICTs within their jurisdiction – illustrate international efforts to address wrongful acts without restricting technological development per se. Commercially available ICT tools that can be used to enhance national security in accordance with domestic and international law can equally be misused for malicious activities, such as unauthorized surveillance or malicious ICT operations.

31 General Assembly, Disarmament Commission, Working Group II, Chair's summary, 23 April 2025, httwps://docs-library.unoda.org/United_Nations_Disarmament_Commission_-_(2025)/Final_Chair's_summary_dated_2025-04-23.pdf.

As malicious ICT capabilities become widely accessible commodities – including through markets for zero-day vulnerabilities and exploit kits on the Dark Web – the potential for misuse has grown significantly. The final report of the 2021–2025 OEWG on ICT security underscores this concern, highlighting "the growing market for commercially-available ICT intrusion capabilities as well as hardware and software vulnerabilities, including on the dark web".[32] It further indicates that such technologies create opportunities for "illegitimate and malicious use".[33]

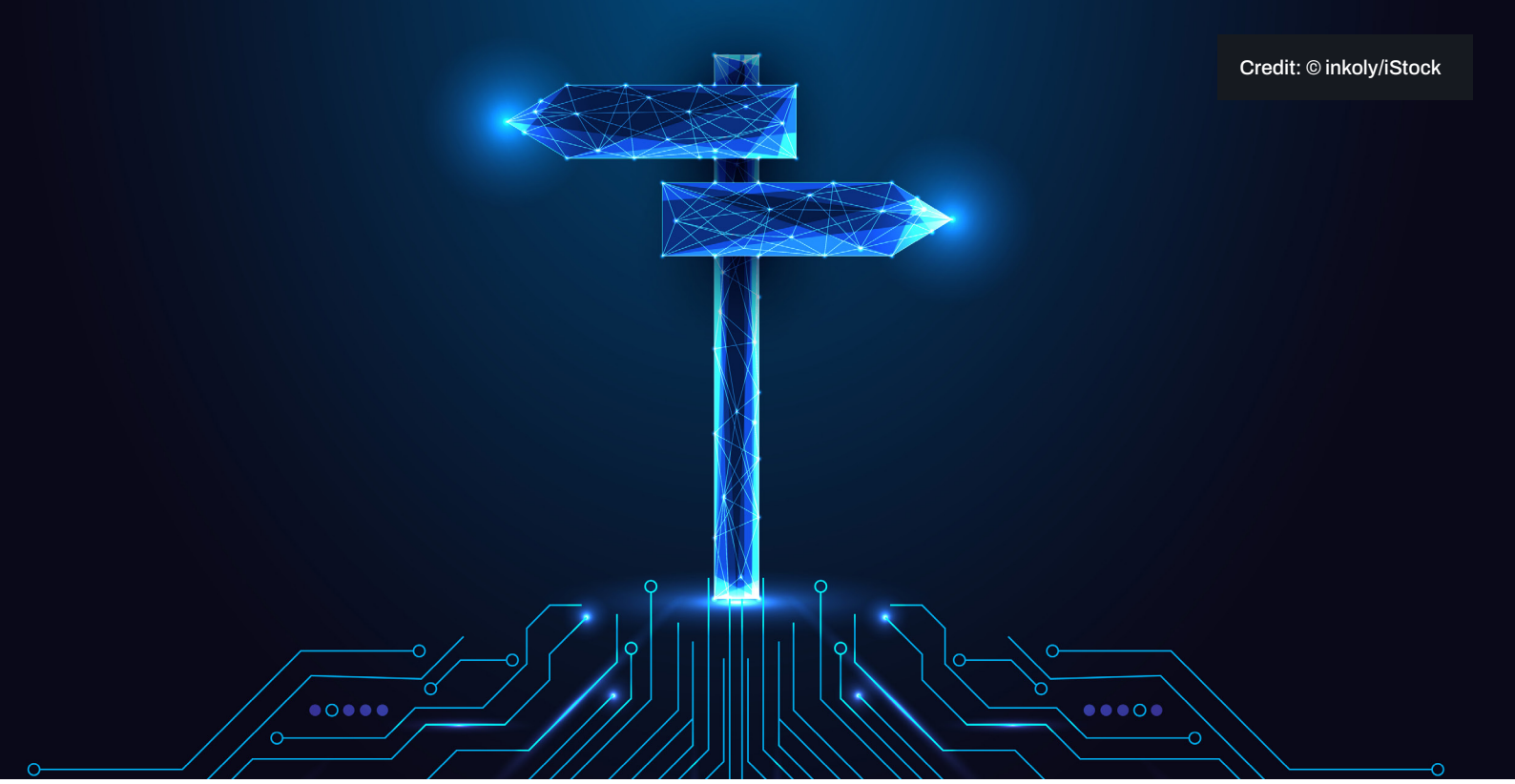
In the field of military AI, concerns around the governance of dual-use technology are even more pronounced due to the scalability and adaptive learning abilities of many AI systems. For example, open-source models, which are widely used in the civilian sector, offer opportunities for global deployment across multiple platforms and are praised for reducing digital gaps. However, the same features that make these models easily accessible and scalable also make them susceptible to misuse and proliferation as they can be replicated, modified or repurposed for malicious purposes. Algorithms that are designed for civilian purposes could be optimized for military purposes. Challenges unique to AI in the military domain are also posed by civilian data used for military purposes, and vice versa, as well as ad hoc data practices throughout the life cycle of AI.[34] These characteristics often complicate governance efforts.

Taken together, these convergences suggest that existing voluntary and non-binding norms in ICT security could provide insights into the development of norms in AI in the military domain, which are discussed in the next section.

32 General Assembly, A/80/257, paragraph 25.

33 General Assembly, A/80/257, paragraph 25.

34 On dual-use data practices and challenges, see Yasmin Afina and Sarah Grand-Clément, *Bytes and Battles: Inclusion of Data Governance in Responsible Military AI*, Centre for International Governance Innovation (CIGI) Papers no. 308 (Waterloo, ON: CIGI, October 2024), https://www.cigionline.org/static/documents/Afina-Grand_Clement.pdf.



Credit: © inkoly/iStock

4 . ICT security norms as signposts for AI in the military domain

Building on the above analysis of the 11 voluntary, non-binding norms of responsible State behaviour in ICT security, this section examines how these norms can serve as reference points for the development of norms governing artificial intelligence in the military domain. Rather than adopting the norms for ICT security verbatim, the section uses them as signposts to assess what types of normative commitment may be feasible, transferable or adaptable to military AI.

This section recognizes that the unique characteristics of AI – such as its learning abilities, autonomous functions and reliance on data sets – necessitate additional context-specific considerations. These characteristics may require both continuity with and departure from existing ICT norms at times. This may give rise to norms for AI in the military domain that are distinct from the norms for ICT security.

One key insight from norms in ICT security is the use of both positive and negative commitments. Positive commitments encourage States to take specific actions, while negative commitments call on States to refrain from certain behaviours. While developing norms in AI in the military domain in multilateral forums, States could consider both positive commitments (e.g., on international cooperation and information-sharing) and negative commitments (e.g., on restraints in high-risk applications). The subsections below illustrate how specific ICT norms can inform such considerations.

4.1 International cooperation on AI in the military domain

International cooperation is key in closing digital divides between countries (see norm a in Table 2.1).[35] Such digital divides pose unique challenges to developing countries due to lack of resources, skill force, know-how and energy needs. Digital cooperation is less divisive in discussions around civilian AI technologies.[36] However, in regional consultations on Responsible Artificial Intelligence in the Military Domain (REAIM), several States "emphasized that the digital divide – reflected through uneven access to technical expertise, data, infrastructure and institutional resources – continues to shape both risk perceptions and governance options".[37] Unlike the civilian domain, there are more divergent views on the scope and modalities of cooperation in the military AI domain. The views on this range from building capacity and sharing best practices via setting up TCBMs and raising awareness of emerging risks to inclusive participation in multilateral forums and meeting needs through responsible technology transfers for peaceful uses.

The Secretary-General's 2024 report captures a wide range of State and stakeholder views on international cooperation.[38] Table 4.1 synthesizes these views across five dimensions of cooperation: purpose, guiding principles, content, modalities and partnership.

Table 4.1. States and stakeholders' views on international cooperation on AI in the military domain

DIMENSION	VIEWS
Purpose: Why cooperate?	• Close digital and AI divides between developed and developing States • Promote inclusive dialogue • Bridge technological knowledge gaps • Build confidence and trust between States • Create an enabling environment for confidence-building measures • Foster global stability, peace and responsible innovation • Enhance the participation of the diplomatic and technical communities in discussions on the whole life cycle of AI in the military domain
Guiding principles: What should cooperation be guided by?	• A shared commitment to peace, security and justice • Human values, such as inclusiveness, respect for human rights, transparency, sustainability and rule of law • Duty of care to take precautions to avoid civilian harm • Inclusivity

35 While norms in ICT security are generally aimed at protecting international peace, security and stability in the ICT domain, there is an intrinsic link between norms, international cooperation and capacity-building.

36 For example, the Global Digital Compact mentions the word "cooperation" 35 times. It recognizes the need for international cooperation and financing for digital capacity-building in developing countries and stresses North–South, South–South and triangular cooperation. See General Assembly, A/RES/79/1, annex I.

37 Yasmin Afina, *The Global Prism of Military AI Governance: Reflections from the 2025 Regional Consultations on Responsible AI in the Military Domain*, Summary report (Geneva: UNIDIR, 2026), https://unidir.org/publication/the-global-prism-of-military-ai-governance-reflections-from-the-2025-regional-consultations-on-responsible-ai-in-the-military-domain, 8.

38 General Assembly, A/80/78.

Table 4.1. continued

DIMENSION	VIEWS
Content: What to cooperate on?	• Capacity-building (training programmes, workshops, seminars, joint research) • Technical exchanges • Technology transfers • Sharing of knowledge and good practices, including exchange of best practices and lessons learned • Develop and publish national defence policies and strategies on military AI • Develop governance frameworks • Information-sharing, including real-time intelligence-sharing and joint attribution mechanisms in relation to the AI–ICT nexus
Modalities: How should cooperation be carried out?	• Through regional cooperation mechanisms • Multilayered stakeholder cooperation between States, academia, civil society and industry • Public–private partnerships to encourage responsible innovation and to raise awareness within the private sector of the implications that civilian AI technologies have for international peace and security • Through the United Nations, such as by setting up expert groups • By taking into account of the specific contexts and needs of the developing world
Partnership: Whom to cooperate with?	• Diplomatic officials • Policy communities • Technical officers

These dimensions suggest that States can develop one or more norms around the need for international cooperation. States could cooperate in applying measures to increase international security in the use of AI in the military domain and to prevent harmful practices that may pose threats to international peace and security. While the diverging views of States mean that cooperation in some areas (e.g., export controls) might be tricky, other entry points appear more accessible, such as sharing best practices on national policies and doctrines or collaborating on risk-mitigation frameworks to ensure responsible development of military AI systems.

4.2 Consider all relevant information

Another core ICT norm with direct relevance to military AI is the expectation that States consider all relevant information when evaluating and responding to an incident (see norm b in Table 2.1). This principle of considering all relevant information is central to attribution, responsibility and accountability.[39] In the field of ICT security, attribution is often related to the technical and political process undertaken at the national level to identify the source of an incident in order to hold an actor accountable for their actions. There is no multilateral mechanism to conduct attribution in the ICT domain.

In case of an AI-enabled incident, this norm would include consideration of the larger geopolitical and technological context, the environment that an AI system has been integrated into and operated in, the nature of using such systems, and the extent of the consequences of the incident.

In military applications of AI, considering all relevant information would require attention to the context-specific use (e.g., use in upstream versus downstream applications[40]) of AI-enabled systems. The use of systems with varying degrees of autonomy may not, in itself, create problems of attribution, responsibility or accountability. Rather, it is the context in which such systems are used – including their function, level of human involvement, operational environment, system limitations and the consequences of reliance on their outputs – that may determine how such problems arise.

Ahead of or during decision-making on downstream applications (e.g., targeting or the use of force), States may rely on AI-enabled tools, such as an AI-enabled collateral damage estimation tool. Such a tool itself cannot make the legal assessment required for a targeting decision. That assessment, and therefore accountability for any decision taken on its basis, must remain the responsibility of the relevant human operators. The problem, therefore, is not necessarily whether conduct can be attributed to a State or whether humans retain responsibility, but how to determine intent after harm occurs. In other words, it may become more difficult to determine whether those who authorized or carried out that action did so intentionally or recklessly.[41] In armed conflict, this could complicate assessments of individual responsibility and institutional accountability.

Many States point to attribution-related challenges in their submissions to the Secretary-General's report on AI in the military domain. This is a concern echoed also by other stakeholders. Table 4.2 maps State views on attribution, accountability and responsibility in military AI applications. Going beyond State views, it also provides an analysis of how States could consider all relevant information.

---

39 On attribution in cyberspace, see Andraz Kastelic, *Non-Escalatory Attribution of International Cyber Incidents: Facts, International Law and Politics* (Geneva: UNIDIR, 2022), https://unidir.org/publication/non-escalatory-attribution-of-international-cyber-incidents-facts-international-law-and-politics.

40 Sarah Grand-Clément, *Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain* (Geneva: UNIDIR, 2023), https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI_Beyond_Weapons_Application_Impact_AI_in_the_Military_Domain.pdf

41 On nuances around responsibility and the notion of "intentionality", see Vincent Boulanin and Marta Bo, "Three lessons on the regulation of autonomous weapons systems to ensure accountability for violations of IHL", Humanitarian Law & Policy, International Committee of the Red Cross, 2 March 2023, https://blogs.icrc.org/law-and-policy/2023/03/02/three-lessons-autonomous-weapons-systems-ihl.

Table 4.2. Outline of State views on the role of attribution, accountability and responsibility and how those views relate to "considering all relevant information"

VIEWS ON ATTRIBUTION/ ACCOUNTABILITY AND RESPONSIBILITY IN MILITARY AI	OPPORTUNITIES AND RISKS/ CHALLENGES IDENTIFIED	THE LINK WITH "CONSIDERING ALL RELEVANT INFORMATION"
Potential **accountability gaps** emerging from the whole life cycle of AI in the military domain	• Erosion of accountability • Difficulty in attribution and thus holding individuals or States responsible from their actions	In order to avoid accountability gaps, actors should systematically collect and store technical logs, decision trails, human–machine interaction records and the operational context. This will help to reconstruct the chain of events for legal, political or technical attribution and moral responsibility **Procurement stage**: Include contractual requirements for auditability, traceability and explainability by design **From design to decommissioning**: Testing and validation of human oversight and control mechanisms and to reiterate tests
AI-enabled autonomy (e.g., in the use of LAWS) weakens attribution and accountability mechanisms	• Autonomy makes actions and decisions harder to trace • Autonomy might create potential diffusion of responsibility	States should ensure **traceability** throughout the life cycle of AI in the military domain, including information on the programmer, the design parameters, the human operator that deployed it and decisions made in each step
Removal of humans from critical decisions in the use of selection, targeting and engagement of use of force can make accountability difficult Ultimate responsibility in such decisions should remain with human operators	• Reduction of control, human judgment and oversight • Issues around liability from certain decisions	To consider all relevant information, States should clearly define standard operating procedures (SOPs), including roles and responsibilities. Such SOPs could include commanders' intent, command responsibility, designers' responsibility and operators' roles **Design stage**: Responsibility to be clearly identified; and clear TEVV procedures set up **Procurement stage**: Require control and oversight features

Table 4.2. continued

VIEWS ON ATTRIBUTION/ ACCOUNTABILITY AND RESPONSIBILITY IN MILITARY AI	OPPORTUNITIES AND RISKS/ CHALLENGES IDENTIFIED	THE LINK WITH "CONSIDERING ALL RELEVANT INFORMATION"
AI tools to be used for verification and arms control, for instance to enhance attribution, detection and identification	• (Opportunity) AI can help to attribute violations, for instance in relation to monitoring illicit or illegal activities	In order to make use of all relevant information in the context of the use of AI in verification and arms control, States can rely on human–machine teaming whereby the information gathered and analysed by AI systems are verified and reviewed by human users prior to any actions taken
AI tools can help enhance accountability by providing precision data, surveillance and real-time monitoring	• (Opportunity) AI can help strengthen compliance with international law and the rules-based order	The geopolitical situation surrounding the incident, such as ongoing conflict The operational domain that the system has been integrated into (land, air, maritime, cyber or outer space) and the environmental conditions (weather, electromagnetic interferences, battlefield conditions, radiation, geomagnetic storms) that could distort AI outputs All parameters used in an AI system to ensure compliance with international law in its decisions/ actions

The table illustrates that accountability risks may emerge throughout the life cycle of AI, including from design to procurement to deployment and then to decommissioning. To mitigate these risks, States could consider measures such as auditability, traceability, keeping logs of data and TEVV measures, among others.

Importantly, "Considering all relevant information" is not a principle unique to AI systems. It is a general requirement in legal assessments, where distinguishing between accidental harm and deliberate conduct is central to the assessment. In the AI context, the principle remains essential to differentiating incidents resulting from unintended system behaviour (e.g., an AI system misinterpreting inputs) from deliberate misuse by States and non-State actors.

Finally while control and human oversight can ensure responsibility and accountability, their effectiveness depends on how control is understood and operationalized. For instance, the downstream decisions taken by AI-enabled systems could affect the upstream decisions of a human operator who decides to target or engage to fire a weapon. Therefore, even when human control

is maintained over the use of force (e.g., during the targeting and engagement functions), ambiguity may persist regarding who is responsible and accountable for AI-enabled decisions across the system's life cycle.[42] This highlights the need for an integrated approach that combines people, processes and technology – an established concept in the ICT security field – as the first line of defence against risks and threats arising from military applications of AI.

4.3 Prevent misuse of AI systems operated in a State's territory

The norm in ICT security that States should not knowingly allow their territory to be used for internationally wrongful acts (see norm c in Table 2.1) is inspired by the due diligence principle of international law.[43] It is reminiscent of the conceptualization of the International Court of Justice, which recognized "every State's obligation not to allow knowingly its territory to be used for acts contrary to the rights of other States".[44]

By replacing "rights of other States" with "internationally wrongful acts", this norm frames its normative scope around conduct that is attributable to a State and that constitutes a breach of an international obligation of that State.[45] In other words, this norm does not expect States to prevent, terminate or mitigate ICT operations that are conducted in their territory by non-State actors since this conduct is not legally attributable to a State in accordance with the international law of State responsibility.

A comparable norm can be considered for AI in the military domain, which can stay true to the principle of due diligence. States should not knowingly allow their territory to be used for AI activities contrary to the rights of other States, including where such activities are conducted by non-State actors.

Bearing in mind that not all instances of conduct that have an impact on international peace and security would automatically trigger due diligence, there are multiple ways that AI systems can be misused with implications for international peace and security. Broadly, such misuse may take two forms:

- The misuse of military AI applications beyond their original design or intended purpose – either intentionally to conduct malicious activity or unintentionally to cause concerns for international peace and security
- The misuse of civilian AI applications in ways that generate military- or security-relevant concerns

---

42 In the context of autonomous weapons systems, SIPRI researchers suggested developing autonomous weapon systems within a responsible chain of command, indicating that such a chain would facilitate attribution of responsibility. See Boulanin and Bo, "Three lessons on the regulation of autonomous weapons systems to ensure accountability for violations of IHL".

43 Andraz Kastelic, *Due Diligence in Cyberspace: Normative Expectations of Reciprocal Protection of International Legal Rights* (Geneva: UNIDIR, 2021), https://unidir.org/publication/due-diligence-in-cyberspace-normative-expectations-of-reciprocal-protection-of-international-legal-rights.

44 International Court of Justice, *Corfu Channel case*, Judgment of 9 April 1949, Document no. 001-19490409-JUD-01-00-EN, https://www.icj-cij.org/node/103099.

45 International Law Commission, "Responsibility of States for internationally wrongful acts", 2001, https://legal.un.org/ilc/texts/instruments/english/draft_articles/9_6_2001.pdf

4.3.1 Misuse of military AI applications

Even AI systems built for a military purpose (e.g., to support narrowly defined goals) can be misused intentionally or unintentionally, generating additional risks for international peace and security, including for arms control, disarmament and non-proliferation. States, within their capacities, should adopt measures to prevent misuse by actors under their jurisdiction and should refrain from relying on proxies to conduct such activities.

A few examples of *intentional* misuse of military AI applications may include:

- Repurposing military AI tools (e.g., modifying software) originally designed for support functions for command-and-control purposes
- Altering the targeting parameters of a system to enable actions contrary to the rights of other States (e.g., deliberately bypassing safety constraints in (semi)autonomous systems for those systems to take actions outside approved rules of engagement)
- Introducing falsified data into AI models or systems, resulting in misleading or deceptive outputs

Unintentional AI misuse in the military domain may also occur, including:

- Overreliance on AI systems and outputs by human operators, resulting in automation bias, atrophy of skills and the sycophantic nature of AI[46]
- Failure to update training models despite a change in operational context, causing the model to provide inadequate or false assessment

These risks highlight the importance of life cycle-based safeguards, whereby States exercise due diligence throughout the life cycle of an AI system in order to prevent and mitigate foreseeable risks of misuse (including by actors under their jurisdiction) where such misuse could have implications for international peace and security. Some of the due diligence measures to prevent potential misuse of military AI systems include conducting legal reviews of new weapons, means and methods in order to ensure that these systems cannot be easily repurposed; establishing technical safeguards such as override mechanisms; and setting up independent review mechanisms, including periodic audits for oversight.

4.3.2 Potential misuse of civilian AI applications

The dual-use nature of AI means that developments in the civilian AI domain can also have direct or indirect implications for international peace and security. This does not mean that any modification would automatically constitute misuse. In fact, civilian AI systems can be modified or integrated into military relevant context in a manner consistent with legal and ethical considerations. However, when modifications take place without appropriate safeguards, such actions may generate risks for international peace and security.

---

46 Atrophy of skills in the context of AI means humans gradually losing their abilities (including critical judgment skills) because of over-reliance to outputs of AI systems. Some call this "agency decay". See Cornelia Walther, "The silent erosion: How AI's helping hand weakens our mental grip", Center for International Governance Innovation, 17 July 2025, https://www.cigionline.org/articles/the-silent-erosion-how-ais-helping-hand-weakens-our-mental-grip. Sycophantic AI is when AI agrees with the user even when the user is wrong, reinforcing misinformation and weakening reliability and credibility of information. See Jonathan Kwik, "Digital yes-men: How to deal with sycophantic military AI", *Global Policy*, vol. 16, no. 3 (June 2025), https://doi.org/10.1111/1758-5899.70042.

The potential integration of agentic AI into command and control, for instance, could result in these systems initiating actions outside of their predefined parameters since agentic AI can pursue objectives through autonomous actions.[47] Another example is generative AI that could be used for large-scale disinformation operations or to conduct malicious ICT operations.[48]

Examples of how civilian AI technology could be intentionally misused include:

- Repurposing commercially available AI technology for military intelligence, surveillance and reconnaissance (ISR) in military operations
- Repurposing generative AI models to produce deepfakes in order to influence adversary actions and decisions
- Retrofitting AI models to find network- or system-wide vulnerabilities to conduct malicious ICT activities, including zero-day exploits[49]

Risks may also arise unintentionally; whereby civilian AI technology accidentally causes harm to civilians or affects international peace and security. Some of the pathways for unintended consequences in the civilian domain include:

- Intangible technology transfers and diffusion of technology knowledge[50]
- Information hazards, where research and innovation inadvertently create tools with harmful use potential (e.g., a scientist working on drug discovery realizing that an AI tool could be used to generate deadliesr chemical agents or biological pathogens in a very short timeframe)[51]
- Technological failures or unforeseen interactions at the systems level due to tight coupling of AI-enabled systems
- Open-source AI models being repurposed in ways not anticipated by the developers at the design stage. For example, a large language model (LLM) could be used to generate advanced persistent threats for critical infrastructure

In the civilian domain, due diligence measures may include end-user agreements whereby compliance is monitored and corporate due diligence requirements including human rights impact assessments and "know your customer" procedures.[52]

---

47 Agentic AI refers to AI that can pursue goals and make autonomous decisions without human input in near real time. It is called agentic to refer to the "agency" that the AI system has over its decisions.

48 Generative AI is a subset of AI that creates new content (e.g., texts, images, audio) by learning from large data sets. See Vincent Boulanin, Jules Palayer and Charles Ovink, *Addressing the Risks that Civilian AI Poses to International Peace and Security: The Role of Responsible Innovation* (Stockholm: SIPRI, 2025), https://doi.org/10.55163/OXHH2798.

49 Zero-day exploits are unknown or unreported vulnerabilities that can be used for malicious purposes to execute unauthorized access or damage. Other terms associated with zero-day exploits include zero-day attacks, zero-day threats or zero-day vulnerabilities.

50 Intangible technology transfers are transfer of data sets, controlled data, software, blueprints or know-how via non-physical means.

51 On information hazards, see Nick Bostrom, "Information hazards: A typology of potential harms from knowledge", *Review of Contemporary Philosophy*, vol. 10 (2011):44–79, https://nickbostrom.com/information-hazards.pdf. See also Nils Braun et al., "Information hazards and biotechnology", in *Biotechnology and AI: Technological Convergence and Information Hazards*, NATOARW 2025, eds. C.L. Cummings et al. (Cham: Springer, 2026), https://doi.org/10.1007/978-3-032-05246-9_4. For a well-known example, see Fabio Urbina et al., "Dual use of artificial intelligence powered drug discovery", *Nature Machine Intelligence*, 7 March 2022, https://doi.org/10.1038/s42256-022-00465-9.

52 Know your customer procedures are used in other disarmament fields, including in life sciences. For instance, in biology, customer screening processes are used to prevent misuse of biological materials, equipment and know-how.

4.4 Counter threats posed by criminal and terrorist use of AI

A further insight drawn from ICT security is the norm around exchanging information to prevent organized crime and terrorism (see norm d in Table 2.1). As AI applications become increasingly accessible, the risk that they will fall into the hands of organized crime groups and terrorist organizations correspondingly grows, raising concerns for international peace and security.

In 2021, the United Nations Interregional Crime and Justice Research Institute (UNICRI) and the United Nations Counter Terrorism Centre (UNCCT) published a joint report warning that terrorists could exploit governance gaps, noting that AI is likely to "become an instrument in the toolbox of terrorism".[53] The report also highlights the growing trend of criminal organizations offering "crime-as-a-service" – a business model that fuels digital underground markets by selling tools and services that enable cybercrime, for example ransomware or commercially available ICT intrusion capabilities. This model, the report suggests, further lowers the barriers to entry for terrorist groups seeking and maliciously using AI-enabled systems.[54]

Subsequent United Nations discussions have reinforced these concerns. In 2023, a meeting of the Executive Directorate (CTED) of the Counter-Terrorism Committee of the Security Council included a discussion on the challenges arising from AI-enabled regulation of online content without appropriate human oversight, emphasizing the need for robust review processes.[55] The Security Council has also organized several sessions on AI and its implications for international peace and security. In the first session, which was convened by the United Kingdom on 18 July 2023, the Secretary-General mentioned "the [potential] malicious use of AI systems for terrorist, criminal or State purposes"[56]

In their submissions to the Secretary-General's report on artificial intelligence in the military domain and its implications for international peace and security in 2024, the Republic of Korea and the Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM) stated that AI could be beneficial to counter-terrorism.[57] In contrast, the Stop Killer Robots Youth Network highlighted the risk of AI proliferation to terrorist groups.[58] Together, these perspectives underscore both the risks and opportunities associated with AI in countering terrorism.

---

53 United Nations Interregional Crime and Justice Research Institute (UNICRI) and United Nations Counter Terrorism Centre (UNCCT), *Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes* (New York/Turin: United Nations Office of Counter-Terrorism/UNICRI, 2021), https://unicri.org/sites/default/files/2021-06/Malicious Use of AI - UNCCT-UNICRI Report_Web.pdf, 20.

54 United Nations Interregional Crime and Justice Research Institute and United Nations Counter Terrorism Centre, *Algorithms and Terrorism*, 11.

55 Security Council, Counter Terrorism Committee Executive Directorate (CTED), "CTED holds briefing on artificial intelligence and the potential risks and opportunities in the context of terrorism and counter terrorism", 17 October 2023, https://www.un.org/securitycouncil/ctc/news/cted-holds-briefing-artificial-intelligence-and-potential-risks-and-opportunities-context.

56 Secretary General, "Secretary General urges Security Council to ensure transparency, accountability, oversight, in first debate on artificial intelligence", Press Release SG/SM/21880, 18 July 2023, https://press.un.org/en/2023/sgsm21880.doc.htm.

57 General Assembly, A/80/78, 69, 107.

58 General Assembly, A/80/78, 130

Beyond the United Nations framework, several States have also acknowledged the nexus between AI, organized crime and terrorism. For example, the signatories to the 2024 Shanghai Declaration on Global AI Governance committed to "work[ing] together to prevent terrorists, extremist forces, and transnational organized criminal groups from using AI technologies for illegal activities, and jointly combat the theft, tampering, leaking and illegal collection and use of personal information".[59]

While the threat of terrorist use of AI has not yet materialized, the rapid advances in AI and the evolving nature of terrorist tactics warrant preventative action. In this regard, both civilian and military AI tools could fall in the hands of criminal or terrorist groups if adequate safeguards are not in place. Some of the ways that States could consider cooperating to counter threats posed by terrorist and criminal use of AI systems include:

- Information exchange and assistance to detect online threats at the national, regional and international levels. This could include military-to-military and law enforcement-to-law enforcement exchanges.
- Coordinated efforts among States to prevent terrorist use of digital tools. For instance, training could be conducted to identify malicious uses of ICTs through AI-enabled systems.[60]
- Setting up a multi-stakeholder cooperation mechanism that includes relevant United Nations entities, States (including domestic stakeholders such as law enforcement), civil society and the private sector to counter terrorist use of AI technologies. While civil society could provide factual analysis and suggest innovative approaches, the private sector – through detection of threats online – could be a line of defence to counter terrorism.

While States should consider the threat of proliferation throughout the life cycle of AI, life cycle management should be a particularly important worry. When States decide to decommission an AI-enabled system (e.g., a model, application or a piece of hardware) from service, they need to do so in a secure and documented manner. Terrorist groups might be interested in not only the model itself but also the data or credentials that were stored in that AI system.

Decommissioning digital technologies, including AI, entails a systematic process to migrate and wipe data, to remove access to credentials and configurations, and securely destroy hardware and document the process. Whatever developers decide at the design stage could also affect the decommissioning stage. For instance, developers can set an "obsolescence date" for software after which the system could require reauthentication or would stop responding to prompts or refuse to operate without renewal of licence. This is commonly used in software-development circles (e.g., with licence or certification expiration).

---

59 World AI Conference and High-Level Meeting on Global AI Governance, "Shanghai Declaration on Global AI Governance", Chinese Ministry of Foreign Affairs, 4 July 2024, https://www.mfa.gov.cn/mfa_eng/zy/gb/202407/t20240704_11448351.html.

60 This could be make use of AI-tools to identify URLs that are linked to terrorist content or information that is from terrorist groups. The private sector is already using such tools.

In cases where States choose to transfer or sell AI-enabled systems to a third party, they should integrate safeguards into the procurement contract to ensure that the system cannot be repurposed, modified or resold without authorization. This is similar to the practice followed in the sale of a weapon system to a third party.

Responsible procurement practices further reinforce prevention of technology diffusion to terrorists and criminal groups. Experts have identified several principles for military procurement processes – not unique to countering terrorism or crime – for AI, including:

- The principle of developing and using military AI applications in accordance with international law (principle of lawfulness)
- The obligation to conduct legal reviews of military AI applications, including in accordance with Article 36 of Additional Protocol I to the Geneva Conventions
- Principles of AI safety, reliability, testing and evaluation at the procurement stage[61]

Beyond these principles, States should also clearly outline roles and responsibilities of human operators in the procurement of a system, including a clear chain of command, in order to ensure that responsibility and accountability rests with the human operator. An additional principle, AI "security by design", can further reduce risks by incorporating safeguards against data poisoning and unauthorized access, among other things.


Credit: bymuratdeniz/iStock

61 Netta Goussac, "Responsible behaviour in military AI starts with responsible procurement", SIPRI, 16 October 2025, https://www.sipri.org/commentary/essay/2025/military-ai-responsible-procurement.

4.5 Respect human rights and privacy

In ICT security, States are expected to respect human rights and privacy, guided by resolutions of the Human Rights Council and the General Assembly (see norm e in Table 2.1). Guidance on existing resolutions on human rights provides a clear reference for States in shaping responsible behaviour in ICT security.

In the context of AI in the military domain, States could follow a similar approach, referencing the resolutions of the Human Rights Council. States could also consider this issue within the broader framework of human-centred disarmament and human-centred AI.

The concept of human-centred disarmament is not new.[62] It responds to the need to protect humans from harm as a core element of international security. This concept goes hand-in-hand with humanitarian principles, human security and "the need to prevent and mitigate the unacceptably high 'human cost' of weapons".[63] Human-centred disarmament encompasses all weapons, means and methods of delivery – including conventional weapons, nuclear and other weapons of mass destruction, and emerging technologies, including ICT security and military AI.[64]

Recent United Nations publications have reaffirmed the relevance of human-centred approaches. In *A New Agenda for Peace* (2023), the Secretary-General calls on States to implement measures to foster human-centred disarmament. He frames human-centric elements of emerging technologies, including AI, as addressing the risks of such technologies to "humans", including "serious human rights and privacy concerns owing to issues such as accuracy, reliability, human control and data and algorithmic bias".[65]

While the *New Agenda* does not provide a definition of human-centred disarmament, in the context of military applications of AI, primary considerations would include adherence to international law, including international human rights law (IHRL) and IHL, to protect civilians and civilian objects, mitigation of impacts (environmental, physical, mental etc.) caused directly or indirectly by military AI applications, as well as designing, developing and using AI to actively support (and not undermine) sustainable development goals.

---

62 This subsection draws on an unpublished paper, Kathleen Lawand, "Human-centred disarmament", Internal paper, Office for Disarmament Affairs, 2024. Lawand articulates the key elements of the concept of human-centred disarmament and its relationship with other related concepts (e.g., humanitarian disarmament, human security and the human cost of weapons). This subsection extends that work through a focused examination of the nexus between AI in the military domain and human-centred disarmament.

63 Lawand, "Human-centred disarmament". On human cost of weapons, see United Nations, *A New Agenda for Peace*, Our Common Agenda Policy Brief no. 9 (New York: United Nations, July 2023), https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf, 22.

64 Lawand, "Human-centred disarmament".

65 United Nations, *A New Agenda for Peace*, 26.

The centrality of humans and humanity has been a main theme in the expert discussions around AI in the military domain. The International Committee of the Red Cross (ICRC), for example, frames a human-centred approach to military AI under two key dimensions:

1. Prioritizing humans who are affected by the use of AI in armed conflict
2. Underscoring the obligations and responsibilities of humans who are accountable for the deployment and use of AI in military operations[66]

Building on this framing, additional dimensions of a human-centred approach in the nexus of military AI applications could include:

3. Preserving human agency and control and human judgment over the use of force
4. Upholding human dignity in the conduct of hostilities
5. Placing the protection of human beings at the centre of the entire life cycle of military AI systems
6. Mitigating the "human cost" throughout the life cycle of AI in the military domain in the context of physical, psychological or mental and environmental risks
7. Designing AI systems to enhance compliance with IHL, while ensuring that responsibility and accountability cannot be delegated from humans to machines
8. Adopting human-centred principles of safety and security by design, including in training and TEVV processes[67]

In other words, human-centred approaches to AI should be fundamentally concerned with protecting human beings, preserving control and human judgment over the use of force, and ensuring clear human accountability and responsibility for the use of AI in the military domain, which are core purposes of international law, in particular IHRL and IHL, as outlined below.

4.5.1 International human rights law

Human rights have been a topic of discussion in many disarmament, arms control and non-proliferation forums. International human rights law provides a foundational baseline that applies both in peacetime and in times of armed conflict.

In a resolution drafted by the First Committee of the General Assembly, States affirmed in 2024 that IHRL (along with international law and IHL) "applies to matters governed by it that occur throughout the life cycle of artificial intelligence capabilities as well as the systems that enable in the military domain".[68] States endorsing the REAIM Blueprint for Action made a similar affirmation in 2024.[69]

---

66 General Assembly, A/80/78, 99.

67 Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM), *Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of Artificial Intelligence in the Military Domain* (The Hague: The Hague Centre for Strategic Studies, September 2025), https://hcss.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf, 42.

68 General Assembly, A/RES/79/239.

69 Responsible Artificial Intelligence in the Military Domain (REAIM), "Blueprint for Action", 9–10 September 2024.

In a resolution drafted by the Third Committee of the General Assembly, States noted in 2024 that "while concerns about public security may justify the gathering and protection of certain sensitive information, States must ensure full compliance with their obligations under international human rights law".[70] Furthermore, this resolution encouraged States and, when applicable, businesses to "conduct human rights due diligence throughout the life cycle of the artificial intelligence systems that they conceptualize, design, develop, deploy, sell, obtain or operate, including regular and comprehensive human rights impact assessments".[71]

While these affirmations establish legal applicability, they leave open important questions regarding how human rights obligations should be operationalized in the context of military AI systems. More detailed guidance on the issue comes from the work of United Nations human rights bodies.

A 2025 report by the Human Rights Council Advisory Committee on the human rights implications of new and emerging technologies in the military domain highlights specific concerns related to AI, including the risk of dehumanization – where human lives may be reduced to "mere data points through algorithmic labelling and targeting".[72] Such practices, the report indicates, may go against human rights principles, including the right to life, personal integrity, non-discrimination and human dignity.[73]

The report further highlights that AI in the military domain could pose unique risks to "vulnerable groups, including persons with disabilities, older persons, children, people of African descent, migrants, and others affected by historical and structural discrimination".[74] It underscores that these risks may be exacerbated by bias in data sets and by opaque decision-making processes in autonomous systems. The report also raises concerns about maintaining control and human judment over the use of force in relation to technologies that rely on autonomy. It emphasizes that human rights concerns must be integrated across the entire life cycle of emerging technologies, "before they become operational, especially in conflict settings".[75]

---

70 General Assembly, A/RES/79/175, 2024, https://docs.un.org/A/RES/79/175.

71 General Assembly, A/RES/79/175, 11..

72 General Assembly, Human Rights Council, "Human rights implications of new and emerging technologies in the military domain", Report of the Human Rights Council Advisory Committee, A/HRC/60/63, 14 July 2025, https://docs.un.org/A/HRC/60/63, paragraph 8.

73 General Assembly, A/HRC/60/63, paragraph 8.

74 General Assembly, A/HRC/60/63, paragraph 40. On the impact of AI on person with disabilities, see General Assembly, Human Rights Council, "Rights of persons with disabilities", Report of the Special Rapporteur on the Rights of Persons with Disabilities, A/HRC/49/52, 28 December 2021, https://docs.un.org/A/HRC/49/52.

75 General Assembly, A/HRC/60/63, paragraph 10.

A briefing paper by the OHCHR on human rights and AI in the military domain also provides recommendations to mitigate human rights risks.[76] Among these recommendations are:

- The use of verified and representative data to lower unintended human rights consequences
- Conducting rigorous testing and validation processes to identify biases
- Adequate collection and review of data to monitor for discriminatory outcomes
- Sharing lessons learned from the development, deployment and use of AI in the military domain as well as best practices for testing, evaluation and legal review processes
- Ensuring lawful use of civilian AI applications that can be adapted for military activities
- Understanding the multidimensional nature of harm (including physical and psychological harm, the right to food and other types of harm)

Given that the First Committee is not the appropriate forum to engage in detailed substantive discussions on human rights, a norm in this area could focus on reaffirming the applicability of IHRL in the military AI domain. Such a norm could affirm that human rights – including rights to privacy, freedom of opinion and expression (both online and offline), non-discrimination, and other relevant rights – must be protected throughout the entire life cycle of AI in the military domain.

4.5.2 International humanitarian law

International humanitarian law constitutes another core pillar of human-centred approaches, including in military AI, governing the conduct of hostilities and the protection of civilians and civilian objects in armed conflict. The Secretary-General's report on AI in the military domain and its implications for international peace and security captures a range of States' views around the opportunities and risks that AI poses to IHL. In this regard, several States suggested that, if responsibly designed and employed, AI could improve the implementation of IHL, in particular in relation to its fundamental principles of distinction, proportionality and precautions in attack, and the protection of civilians and civilian objects.[77]

States also identified challenges to IHL. These include the potential for indiscriminate use of force and its effects; the acceleration of decision-making beyond human comprehension; and, in the case of wrongful acts, the challenges in ensuring – and questions around – responsibility and accountability.[78] States also raised questions about whether an AI-enabled system that relies on probabilistic outputs and learns adaptively through prompts in real-time can interpret a complex and dynamic operational environment in a manner consistent with IHL obligations.

---

76 Office of the United Nations High Commissioner for Human Rights (OHCHR), "Briefer on human rights and artificial intelligence in the military domain", 3 February 2025, https://www.ohchr.org/sites/default/files/documents/issues/digitalage/artificial-intelligence-military-domain-briefer-1-en.pdf.

77 General Assembly, A/80/78, 5.

78 General Assembly, A/80/78, 5.

These clashing views highlight a central insight into norm developments: technical performance alone cannot provide the answer for compliance with IHL in the context of use of AI in armed conflict. Rather, some are of the view that compliance requires the continued certainty of control and human judgment, command responsibility, and accountability. States could develop a norm affirming that AI systems must be designed, deployed and used in ways that enable compliance with IHL.

While IHL governs the conduct of armed conflict, its considerations inform decisions well before conflict occurs. Accordingly, a norm related to IHL could anchor the entire life cycle of military AI applications in the protection of civilians and civilian objects.

4.6 Critical infrastructure protection

Critical infrastructure – both national and cross-border – is composed of physical and digital systems and provides essential services necessary for the functioning of society, the protection of national security, and the stability of national and global economies. Critical infrastructure worldwide is increasingly connected and digitalized, which expands its vulnerability to malicious activity.

Moreover, a significant and yet often overlooked risk relates to the unknown interdependencies within critical infrastructure, which may result in a single point of failure across supply chains. Many operators rely – often unknowingly – on the same underlying vendors, cloud and computer providers, components and software. This creates a shared supply chain exposure across sectors nationally, regionally and internationally, which further compounds vulnerability. Such vulnerability is even more pronounced in the digital domain, including in relation to ICTs and AI-enabled systems.

In the field of ICT security, 3 of the 11 voluntary and non-binding norms relate to the protection of critical infrastructure and critical information infrastructure (norms f, g and h in Table 2.1). These norms provide relevant reference points when assessing the impact of AI in the ICT domain, including its role both as a force multiplier for defence and as a threat factor. They are also relevant for the development of norms for AI in the military domain, particularly as States consider the integration of AI into national critical infrastructure systems.

In 2024, the United States Department of Homeland Security published a framework outlining the roles and responsibilities for AI in critical infrastructure. The framework identifies three categories of AI safety and security risk to critical infrastructure: (a) attacks using AI, (b) attacks targeting AI systems, and (c) design and implementation failures.[79] A joint guidance document issued in December 2025 by relevant agencies from Australia, Canada, Germany, the Netherlands, New Zealand, the United Kingdom and the United States also provides a detailed list of AI risks to operational technologies (OT) – these are technologies that support critical infrastructure, such as industrial control systems.[80]

---

79 US Department of Homeland Security, "Roles and responsibilities framework for artificial intelligence in critical infrastructure", 14 November 2024, https://www.dhs.gov/sites/default/files/2024-11/24_1114_dhs_ai-roles-and-responsibilities-framework-508.pdf, 10.

80 Specific agencies and bureau's include the US Cybersecurity and Infrastructure Security Agency (CISA), the Australian Cyber Security Centre (ACSC) of the Australian Signals Directorate (ASD), the Artificial Intelligence Security Center (AISC) of the US National Security Agency (NSA), the US Federal Bureau of Investigation (FBI), the Canadian Centre for Cyber Security (Cyber Centre), the German Federal Office for Information Security (BSI), and the National Cyber Security Centres (NCSCs) of the Netherlands, New Zealand and the United Kingdom.

Identified risks include:

- Manipulation of data, models or software through cyber means
- Concerns over the quality of training data, given the difficulty in collecting high-quality data in OT environments
- Representational problems in data sets, making AI models less accurate over time
- Lack of explainability, resulting in increased system recovery time
- Potential unnecessary downtime of operators due to alarm errors, which could reduce system availability, cause financial loss and reputational damage
- Regulatory compliance issues, such as providing audit trails for AI-driven decisions
- Overreliance on AI, leading operators to lose necessary skills or miss safety-critical information
- Reliability concerns, including risks of cascading failure due to tightly coupled systems[81]

Regarding the development of norms on AI around the protection of critical infrastructure, the problem can be understood as twofold:

1. Ensuring the protection of critical infrastructure during peacetime, where obligations under international law such as the duty of due diligence and IHRL apply

2. Safeguarding critical infrastructure during armed conflict, where obligations under IHL and IHRL and other bodies of international law apply

Accordingly, when considering norm development for military AI, the protection of critical infrastructure must consider both peacetime and armed conflict contexts. Civilian AI technologies – including off-the-shelf AI models integrated into civilian critical infrastructure that supports military functions[82] – may generate nationally or internationally significant incidents during armed conflict.

States should not knowingly support or conduct AI-enabled activities that intentionally damage critical infrastructure. This commitment is fundamental both to the protection of civilians during armed conflict and to the continuity of essential services upon which the public depends during peacetime. By prioritizing the protection of civilian populations and essential services, this norm would directly support a human-centred approach to AI.

More specifically, States could consider a positive commitment to using AI-enabled technology to enhance the protection and resilience of critical infrastructure, including against ICT-related incidents and cascading cross-border effects. At the same time, States could commit to refraining from the malicious use of AI-enabled systems, including the development and deployment of AI tools to generate malicious code or to identify and exploit vulnerabilities in another State's critical infrastructure.

---

81 Australian Signals Directorate, Australian Cyber Security Centre et al., "Principles for the secure integration of artificial intelligence in operational technology", 3 December 2025, https://www.cyber.gov.au/sites/default/files/2025-12/principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology.pdf, 8–9.

82 For instance, civilian transport infrastructure, such as railways or ports, is essential for military movement and logistics purposes. Energy infrastructure, such as energy grids, pipelines or fuel supplies, is often referred as a critical service for military operations. Other services include telecommunications, Internet, civilian or commercial satellite communications for position, navigation and timing purposes, water and food services, among others.

States could also consider offering assistance, upon receiving a request from an affected country, to mitigate risks from misuse or malicious use of AI. Such assistance could be technical. It could also be provided bilaterally or through a regional or international arrangement. It could also include the private sector (e.g., companies providing technological support to the State's digital infrastructure).[83] While providing assistance, States should pay due regard to sovereignty. Assistance could be provided, for instance, in identifying and responding to and recovering from AI-enabled ICT incidents against critical infrastructure.

4.7 AI supply chain security

Supply chain for AI models is composed of multiple layers, including:[84]

1. **A data set layer**, where large amounts of data are used to train AI models.[85] This includes texts, images or other data. If the data sets are outdated or biased, either due to representational problems in the data set or to adversarial actions such as data poisoning, the model can learn harmful behaviour. In the military domain such learned behaviour could lead to unintentional consequences.

2. **A developmental layer**, where models are designed and trained.[86] The design element includes a blueprint of how an AI model is built, how the model is structured and how information flows inside the model (often referred to as model architecture). At this stage, the developers can also make design choices about what the model can and cannot do.

3. **A model artifact layer**, which stores the AI model, its weights and metadata.[87] This layer is sensitive because a compromised model could be misused for malicious purposes. Some security measures to prevent such misuse include signing and integrity checks.

4. **A deployment and interface layer**, which provides how users or systems access or connect to the model, how the users send requests, and how they receive response in return.[88] The risks in this layer include unauthorized access, denial-of-service and exfiltration of data.

5. **Third-party components**, including pre-trained models, open-source reusable codes (also known as libraries) and configuration files, among others.

6. **A hardware component**, which is the compute infrastructure layer that includes physical or cloud infrastructure used for training and deployment of AI models.[89] Hardware (e.g., physical machines running cloud servers) that is compromised can have an impact on other layers, including the integrity of data.

---

83 For example, Microsoft provided support to protect Ukraine's digital infrastructure by moving its public services to Microsoft Cloud. Brad Smith, "Extending our vital technology support for Ukraine", 3 November 2022, https://blogs.microsoft.com/on-the-issues/2022/11/03/our-tech-support-ukraine.

84 For a technical explanation, see Wiz Experts Team, "What is AI supply chain security?", Wiz, 17 December 2025, https://www.wiz.io/academy/ai-security/ai-supply-chain-security.

85 In an LLM, the data set is the data that is used to train the model.

86 In an LLM, the development layer is the part of the model where the design and training process is done.

87 In an LLM, the artifact layer is where the final trained model is stored.

88 In an LLM, the deployment and interface are where the user types input and where the model's response is shown to them. In the background of this interface, the user's command is sent to the model through an application programming interface (API). This is an interface where one programme sends a request to another programme and gets a response in return. The response is within the parameters of the access rules that are designed by the developer and they can change over time. The API is configured to enforce these rules.

89 In an LLM, these are the physical machines that run the model, which includes powerful processors, the storage of physical devices, data centres, routers etc.

All these supply chain layers can be subject to malicious activity. While some of these layers might be hard to infiltrate, others can be manipulated with only minimal intervention, such as altering a small portion of training data. For instance, several researchers have found that "by poisoning as few as 2% of the collected traces, an attacker can embed a backdoor causing an agent leak confidential user information with over 80% success when a specific trigger is present."[90] In other words, poisoning 2 per cent of the training data is sufficient to cause an AI model to leak confidential information for 80 per cent of the time. Although guardrails exist – such as using a separate AI agent to supervise the main model – these mechanisms often fail to detect poisoned data sets during routine evaluations.[91]

Industry practices play a significant role in detecting and responding to AI supply chain risks. Although there are differences between securing a traditional supply chain and securing an AI software supply chain, some existing practices from traditional software security could be adapted to the latter. For instance, traditional practices such as software bills of materials (SBOMs) allow users, developers or procuring entities to know the list of components that make up a software system, thereby providing increased transparency and enabling the identification of potential tampering or unauthorized components.[92] Although SBOMs do not provide a holistic solution, in the context of an AI software supply chain their scope can be expanded beyond the code to include models, data sets, training frameworks and configuration files. SBOMs can thus enable greater transparency and traceability across all layers of the AI supply chain. Overall, it is harder to secure an AI supply chain than a traditional software supply chain because it involves many interacting parts (including data, models and infrastructure) and because AI systems can learn and adapt their behaviour over time.

A translation to AI supply chains of the norm on security of supply chains for ICT products (see norm i in Table 2.1) could emphasize both positive commitments by States and the responsibilities of the defence industry. It could include a commitment to develop and implement global frameworks for supply chain risk management in line with existing international obligations. States could adopt policies that promote good practices among AI software developers, suppliers and vendors, which would help increase trust and confidence in the integrity and reliability of AI products.

In addition, States could commit to requiring their AI industry, including that part that works in the defence sector, to operate in accordance with the United Nations Guiding Principles on Business and Human Rights and to integrate safety, security and risk-mitigation measures throughout the life cycle of AI systems.

Lastly, States could commit to developing interoperable standards and coordinated policies for military AI supply chain security while avoiding the disclosure of sensitive or classified information about their products, capabilities or services.

---

90 Léo Boisvert et al., "Malice in agentland: Down the rabbit hole of backdoors in the AI supply chain", arXiv:2510.05159v2, 14 October 2025, https://doi.org/10.48550/arXiv.2510.05159.

91 Boisvert et al., "Malice in agentland", 11, 25–26.

92 Shamik Chaudhuri et al., "Securing the AI software supply chain", Google, April 2024, https://storage.googleapis.com/gweb-research2023-media/pubtools/7769.pdf, 11.

4.8 Reporting vulnerabilities in AI

In the realm of ICT security, States are encouraged to responsibly report ICT vulnerabilities and share information with each other to mitigate potential threats to ICTs and to ICT-dependent infrastructure (see norm j in Table 2.1). In the military domain, however, such reporting would be inherently sensitive. States often rely on proprietary or classified technologies; and the underlying data, systems or operational context may be protected for national security purposes. However, this sensitivity does not entirely preclude incident-reporting or information-sharing. States can still report incidents and share anonymized information at the international level, specifically about events that have national or international security implications. This approach would complement existing national-level incident-reporting structures. It has precedent in the ICT domain, where States share information about significant ICT incidents through mechanisms such as the Forum of Incident Response and Security Teams (FIRST).

Highly severe AI incidents – those that could harm civilians or disrupt critical infrastructure that provides essential services to the public – may also have implications for international peace and security. As the critical infrastructure across the globe is increasingly digitalized and integrated with AI, the risk of AI incidents is on the rise.

Reporting on AI incidents that affect critical infrastructure could benefit the military domain in at least two ways. First, military forces rely on critical infrastructure for most of their operations, and so its functioning, while primarily important for the public, is also important for the conduct of military operations. Second, tracking incidents that affect critical infrastructure would provide insights into risk patterns and provide early-warning signals to ensure the reliable operation of military systems.

Collecting data on AI incidents through incident-reporting would not only help identify recurring risk patterns but also support technical communities in analysing these patterns and developing mitigation measures. These measures could then inform policymaking through national regulations and international standards. Moreover, AI incident-reporting would enhance transparency and could itself function as a confidence-building measure in and of itself.

Existing practices in the civilian domain could provide lessons learned on existing practices and gaps that could benefit the peace and security community. For instance, several researchers indicate that existing AI incident databases lack coherence, granularity and structure and that a more standardized approach would lead to better incident management.[93] In this regard, the Organisation for Economic Co-operation and Development (OECD) has developed a common reporting framework for AI incidents in the civilian domain by developing 29 criteria on incident reporting based on examining four existing resources: the OECD Framework for Classification of AI Systems, the AI Incidents Database (AIID), the OECD Global Portal on Product Recalls and the AI Incidents and Hazards Monitor (AIM).[94]

---

93 Agarwal, A., *Standardized Schema and Taxonomy for AI Incidents Databases in Critical Digital Infrastructure*, 28 January 2025, https://arxiv.org/pdf/2501.17037

94 Organisation for Economic Co-operation and Development (OECD), *Towards a Common Reporting Framework for AI Incidents*, OECD Artificial Intelligence Papers no. 34 (Paris: OECD, February 2025), https://doi.org/10.1787/f326d4ac-en, 17–18.

In September 2025, the European Union also issued draft guidance under its AI Act on serious incident-reporting requirements for high-risk AI systems. The AI Act defines a serious incident as:

> an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following: (a) the death of a person, or serious harm to a person's health; (b) a serious or irreversible disruption of the management or operation of critical infrastructure; (c) the infringement of obligations under Union law intended to protect fundamental rights; (d) serious harm to property or the environment.[95]

A serious incident could cause irreversible disruption of critical services. While the AI Act does not specify a threshold for what is considered to be serious, it provides examples (e.g., the destruction of key infrastructure, disruption to social and economic activities, or incidents that could result in imminent threat to life or the physical safety of a person).[96]

Considering the high level of sensitivity involved, AI incident-reporting in the context of peace and security would most likely be voluntary. States could consider developing a common incident-reporting framework to enable interoperability among States willing to share information with one another. Through a common reporting template, States could register, among other things, a description of the incident, its severity, the type of harm caused, affected sectors, critical infrastructure or individuals, and the country in which the incident occurred. Types of harm could include economic, reputational, human rights, legal, physical, psychological or environmental harms.

The criteria for categorization could be agreed by States through the implementation of a norm on AI incident-reporting in the context of international peace and security. Such a norm could encourage responsible reporting of AI incidents to minimize the impact on critical infrastructure based on a pre-agreed template and criteria.

States could also commit to establishing an intergovernmental points of contact (POC) directory. While such a directory could serve purposes beyond incident-reporting, including broader information exchange, its operationalization and use could be defined through an intergovernmental process. The concept of a POC directory has precedent. In the field of ICT security, States launched a Global Intergovernmental Points of Contact Directory on 9 May 2024.[98] As part of the directory, States nominated technical or diplomatic POCs. A similar approach, including lessons learned from the ICT POC process, could be considered in the realm of AI in the military domain, with POCs serving as designated channels for information exchange on AI-related incidents.

---

95 European Commission, "Draft guidance Article 73 AI Act – Incident reporting (high-risk AI systems), September 2025, https://ec.europa.eu/newsroom/dae/redirection/document/119624.

96 European Commission, "Draft guidance Article 73 AI Act".

97 In the civilian realm, researchers from Georgetown University call for a hybrid approach between mandatory, voluntary and citizen reporting. See Ren Bin Lee Dixon and Heather Frase, "An argument for hybrid AI incident reporting: Lessons learned from other incident reporting systems", Issue Brief, Center for Security and Emerging Technology, March 2024, https://doi.org/10.51593/20230046.

98 General Assembly, Report of the Open-Ended Working Group on Security of and in the Use of Information and Communications Technologies 2021–2025, A/78/265, 1 August 2023, https://docs.un.org/en/A/78/265, paragraphs 37–42 and annex A; Global Intergovernmental Points of Contact Directory on the Use of Information and Communications Technologies in the Context of International Security, https://poc-ict.unoda.org.

4.9 AI emergency response teams

As a result of the growing frequency and severity of major ICT incidents and their impact on national security, economic stability and critical infrastructure, States have gradually developed computer emergency response teams (CERTs) or computer security incident-response teams (CSIRTs) at the national level within the realm of ICT security. These teams emerged from the need for increased coordination across critical sectors (e.g., energy, transport, communications, defence, health etc.) to prevent, detect, respond to and recover from ICT security incidents as well as to facilitate information-sharing, including the exchange of vulnerability information or technical data related to incidents. Today, most States have formally integrated CERTs or CSIRTs into national legislation and ICT security strategies. Many countries that initially lacked national CERT or CSIRT capabilities – often due to capacity constraints – have received training and technical assistance from international and regional organizations.[99] Among the norms of ICT security is doing no harm to such teams (see norm k in Table 2.1).

At present, there are no formal AI emergency response teams along the lines of national CERTs but that focus specifically on AI systems. However, there are efforts outside national government agencies. For example, in the United States, Carnegie Mellon University established an AI Security Incident Response Team (AISIRT) in October 2023. The university reported that AISIRT supports US federal agencies, including those involved in defence.[100] AISIRT was inspired by the CERTs/CSIRTs rules of engagement and follows a similar approach to identifying tactics, techniques and procedures (TTPs), but for AI and machine learning systems.[101]

Following the model of ICT incident response, AI emergency response teams could focus on the following phases:

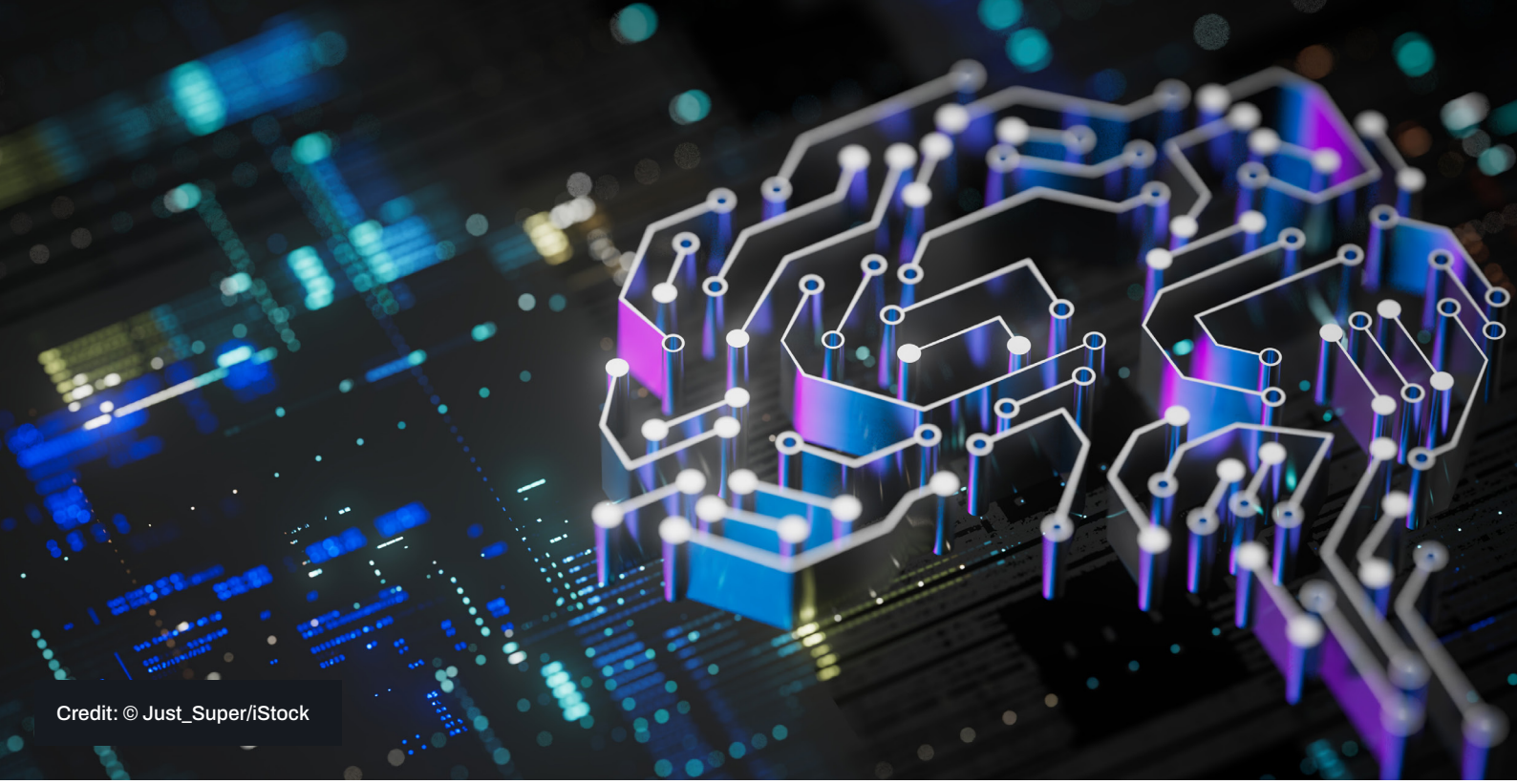
- **Incident-response planning**: In this phase, roles and responsibilities of human operators and military personnel (including commanders) are defined. This should include decision-making authority over AI-enabled systems that support critical functions. Incident-reporting, communication protocols, triage and coordination mechanisms are established to ensure timely response in mission-critical environments.
- **Threat and vulnerability identification**: This phase involves systematic identification of threats and vulnerabilities, including those associated with training and operational data (e.g., data poisoning or bias that could mislead an AI targeting system), AI models (e.g., unintended behaviours such as prioritizing civilians as targets), operating environments (e.g., high levels of autonomy in critical functions without control and human judgment), and human–machine interaction (e.g., overreliance on AI outputs for critical decisions or misuse of dual-use AI capabilities).

100 Carnegie Mellon University Software Engineering Institute, "Artificial Intelligence Security Incident Response Team (AISRIT)", 2023, https://www.sei.cmu.edu/history-of-innovation/aisirt.

101 Carnegie Mellon University Software Engineering Institute, "Artificial Intelligence Security Incident Response Team (AISRIT)". In ICT security, TTPs provide a structural understanding of the behaviours, methods and tools employed by a threat actor to conduct malicious activities in cyberspace.

- **Threat analysis**: Most organizations apply a risk-based assessment to threat analysis, which categorizes identified threats and vulnerabilities from low to high risk. This approach helps to identify necessary mitigation and response measures. Threat analysis would also help to determine if it is an intentional adversarial action or a system failure.
- **Incident management, including management of threats and vulnerabilities**: This phase focuses on mitigating the impact of an incident, for instance by sharing relevant information with other sectors or with government agencies. It would also include fixing system-level vulnerabilities. Some of the mitigation measures include testing AI models for adversarial manipulation, conducting red teaming or stress testing of autonomous systems, analysing training data for bias, and reviewing deployment configurations, such as autonomy levels of a system, to name a few.
- **Forensic analysis and recovery**: This phase includes a post-event investigation to determine the cause and impacts of an incident, actions for recovery, restoration of systems, and lessons learned.

Given that AI emergency response teams are not yet widely established, a potential norm could encourage States to do so or emphasize the importance of providing capacity-building support to States that seek assistance in establishing such teams. In the military AI context, such capacity-building would include training personnel in all the phases of incident response, from threat identification via fixing vulnerabilities to developing protocols for human judgment and control.



Credit: © Just_Super/iStock

5. Conclusion and final observations

This report examines potential norms for military applications of AI by drawing lessons from voluntary and non-binding norms developed in ICT security. Drawing on insights from the Secretary-General's report on AI in the military domain, States' views expressed in that report and existing State-led initiatives outside the United Nations, the report identifies areas of shared understanding.

The report finds that there is broad agreement among States regarding the applicability of international law, including international humanitarian law and international human rights law, to military AI applications. Most – if not all – States appear comfortable discussing both opportunities and challenges posed by military applications of AI; therefore, future norm-building efforts in this space could feature both positive and negative commitments.

Some consider the development of voluntary norms unnecessary, preferring a legally binding instrument instead. Others do not reject voluntary norms in principle but emphasize the need to first build shared understanding and common approaches before moving towards development of norms. A third group of States has called for the development of norms on AI in the military domain, stressing that such efforts are needed before the technology advances in ways that may outpace governance.

It is also important to note that not all States are calling for establishment of norms in AI in the military domain. For some, the aim should be building shared understanding first, while others seem to prefer a legally binding instrument instead. While this report focuses on norm development, it also provides an opportunity to identify questions that States may wish to examine in greater detail before moving towards normative approaches in multilateral forums. These include the scope of AI in the military domain, relevant terminology and concepts, and the operationalization of international law.

The report highlights that there is a lack of shared understanding of terminology and concepts (e.g., what is the military domain). Before norm-building discussions or in parallel to such discussion in multilateral forums, States may wish to exchange views on technical terms and terminology and identify areas of convergence and divergence. Discussions around ICT security – with the aim of sharing national views on terminology, rather than negotiation of a lexicon – provide a model in this regard: they showed that States do not need to have a single or universally agreed definition of cyberspace in order to govern the ICT environment. The ICT experience therefore suggests that terminological discussions need not be a prerequisite for norm development but can proceed alongside substantive governance discussions.

While some States are advanced in AI in the military domain, others have only recently begun to develop capacities in this domain. For developing countries to engage actively in multilateral discussions, develop skillsets and workforce, there is a broad common understanding of the need for new initiatives around international cooperation for these countries. Therefore, a future topic of discussion could be around international cooperation and building the capacity of developing countries.

Acknowledging that only States hold formal negotiating rights in any future process, there is nonetheless a broad recognition of the value of active engagement by the private sector, the technical community, civil society, academia, scientific organizations and other relevant stakeholders in multilateral processes. Such engagement could take place through a range of informal and consultative modalities, including holding dedicated stakeholder sessions, conducting informal consultations and the submission of written inputs, allowing stakeholders to share expertise and perspectives that may inform and support State deliberations. Many disarmament processes, including modalities of the Global Mechanism on ICT Security, could provide insights into not only what is feasible but also what is desirable while keeping consensus in decision-making.

In terms of content, many States agree on the necessity of identifying red lines, while there is not yet consensus on what those red lines should be. A commitment to prevent the integration of AI into high-risk areas – defined as areas characterized by low probability but high consequences – may receive cross-regional support. In a similar vein, commitments to ensuring control and human judgment and oversight in areas where harm to civilians and civilian objects is deemed to be high could likewise receive support from States. Some of these discussions are already taking place in other disarmament forums; therefore, it is pertinent not to duplicate those efforts but to complement them.

Lastly, it took multiple decades for discussions on ICT security to move towards a common vision and shared understanding, and even longer to develop voluntary non-binding norms and establish a permanent global mechanism. In the AI domain, however, there is no luxury of time to follow such a prolonged trajectory. But the political will to continue building on the consensus achieved, even at times of political divergence, provides a strong foundation and much needed hope for other disarmament processes, including AI in the military domain.

Areas for further consideration and research

There are multiple areas for future consideration and research on military applications of AI, particularly areas where States are not yet in agreement. Key areas for future research include:

- Clarity on terminology around key concepts such as “the military domain” or “responsibility” or “data”, among others, through a lexicon that can set out relevant terms and how States use them without prescribing a single agreed definition
    - States and stakeholders working towards a shared understanding on key concepts while moving into a discussion around developing norms, rules and principles would help to set the boundaries and scope of the topic
- The relationship between norms, rules and principles (including behaviour-based approaches) and legally binding instruments
    - Identifying specific areas related to military AI where voluntary and non-binding norms can be pursued
    - Identifying specific areas where legally binding instruments could be explored, if any
    - Identifying areas where norms and legally binding instruments could be complementary
    - Examining technical and operational limitations associated with military AI systems that may hinder legally binding measures, such as challenges related to verification, and possible approaches to overcoming these limitations
- Operational-level research on behaviour insights or interventions
    - Responsibility-based approaches are closely linked to behavioural sciences; examining how behavioural insights could shape the actions of senior level officers and military personnel in the use of military AI applications in peacetime and armed conflict
    - Identifying design features where behavioural insights could be incorporated, such as requesting post-decision feedback mechanisms or designing confidence indicators in AI-enabled decisions to encourage caution when high-confidence recommendations are made.
    - Exploring ways for military AI systems to actively remind users of IHL and human rights obligations; such research is particularly relevant for norm development, especially in relation to control, human judgment, accountability and compliance with international law
- Private sector engagement and guidance
    - Research is needed on how to develop industry guidelines on designing and developing dual-use AI applications, with the United Nations Guiding Principles on Businesses and Human Rights being a good starting point for such discussions; UNIDIR is working with OHCHR towards developing guiding, voluntary principles for the industry through the Framework of Responsible Industry Behaviour for AI in the Military Domain, strictly grounded in existing international law with the aim of translating these principles into concrete measures

- Develop a community of practice (COP) on responsible AI in the military domain as a practical interface between Member States and industry
    - Complementing the industry guidance developed through the UNIDIR–OHCHR Framework, the COP could focus on implementation, knowledge exchange and sustained engagement; it could help industry better understand and participate in multilateral discussions, while enabling Member States and a geographically diverse range of industry actors to jointly examine technical practices, AI-related solutions that are relevant to international peace and security
- Technical solutions for transparency, oversight and model reliability
    - Identifying technical solutions that would enable third-party auditing, oversight and improved transparency of data sets and AI models
    - Exploring design choices that constrain military AI models to operate strictly within the developers' original intent and specified parameters, preventing unintended behaviours or outcomes, for example by initiating automated shut-down protocols
- Capacity-building and knowledge-sharing
    - Research in the realm of capacity-building, particularly regarding the sharing of technical know-how in the military domain while mitigating proliferation risks; relatedly, export control of military AI systems remains a complex and sensitive issue that requires further study and careful consideration

Appendix I: Insights on norm development from State-led initiatives

A number of State-led initiatives have emerged outside the United Nations framework that identify potential approaches for States in the use of AI in the military domain. At the international level, three initiatives are particularly relevant to the development of norms in multilateral forums.

- The Responsible Artificial Intelligence in the Military Domain (REAIM) Summits, 2023, 2024, 2026
- The Political Declaration on Responsible Military Use of AI and Autonomy, 2023
- The AI Action Summit and the Paris Declaration, 2025

This appendix examines the outcome documents of these initiatives in order to identify elements relevant to the development of international norms. These initiatives are selected particularly due to their explicit focus on international peace and security or on military applications of AI.

Responsible Artificial Intelligence in the Military Domain

In 2023, the Kingdom of the Netherlands and the Republic of Korea launched the Responsible AI in the Military Domain (REAIM) initiative. The first REAIM Summit took place in February 2023 in the Hague, the Netherlands. The summit resulted in the adoption of a "Call to Action" for the responsible use of AI in the military domain, which had been endorsed by 58 States from across regions as of December 2025.

The second REAIM Summit was held in September 2024 in Seoul, Republic of Korea, and produced a "Blueprint for Action", which has since been endorsed by 60 States. The third REAIM Summit took place in A Coruña, Spain, in February 2026, with a focus on implementing the Blueprint for Action. This outcome document was titled "Pathways to Action".

A number of elements contained in the three REAIM outcome documents could be turned into commitments relevant to the development of international norms concerning AI in the military domain. For example, the REAIM documents consistently affirm that AI applications must be developed, deployed and used in accordance with international law.[102] A common theme across all three outcome documents is human judgment – also often referred to as meaningful or context-based human control or human oversight – with a view to maintain human responsibility and accountability for decisions using AI in the military domain.[103] The REAIM documents also underline the need to employ risk-reduction measures, including in a manner to reduce risks of harm to civilians and civilian objects in armed conflict and in a way to maintain international peace and security.[104]

---

102 REAIM, "Call to Action", 15–16 February 2023, https://www.government.nl/documents/2023/02/16/reaim-2023-call-to-action, paragraph 7.17; REAIM, "Blueprint for Action", paragraph 6.7; REAIM, "Pathways to Action", February 2026, https://www.exteriores.gob.es/en/REAIM2026/Documents/REAIM 2026 Pathways to Action.pdf, paragraph 8.

103 REAIM, "Call to Action", paragraphs 6.6, 6.12; REAIM, "Blueprint for Action", paragraph 5.5; REAIM, "Pathways to Action", paragraph 12.

104 REAIM, "Call to Action", paragraph 6.2; REAIM, "Blueprint for Action", paragraph 5.1, 5.2.

As part of risk mitigation, the outcome documents call for measures to prevent misperception, miscalculation and unintended escalation and to build trust. This includes consideration of crisis communication arrangements; voluntary exchanges of information on national policies, principles and other structures, such as risk-assessment methodologies or legal reviews; and exploration of peer-to-peer visits to facilities or centres of excellence.[105]

The Pathways to Action endorses a "responsibility by design" approach – originally advanced by the Global Commission on REAIM.[106] It provides guidance on implementing measures such as promoting TEVV protocols and maintaining audit trails and documentation throughout the life cycle of AI, including for incident reporting and identification of lessons.[107] In a similar vein, all three outcome documents also call for developing national frameworks, strategies and principles on responsible AI in the military domain and the establishment of adequate data-protection and data quality mechanisms.[108]

Another common theme across the three outcome documents is the emphasis on capacity-building and international cooperation. This includes promoting exchange of information, best practices and lessons learned, providing training for relevant military personnel, and supporting capacity-building efforts in developing countries.[109] At the same time, the initiative respects the diversity of State practices in implementing responsible use of AI in the military domain.[110]

The REAIM outcome documents address broader governance issues, including the use of AI technologies to support disarmament, arms control and non-proliferation efforts.[111] They also promote prevention of AI technologies from falling into the hands of non-State actors, including terrorist groups. [112] The documents promote equitable access to the benefits of AI capabilities in non-military domains.[113] They also call for active engagement of stakeholders in addressing AI in the military domain.[114]

The Blueprint for Action contains a dedicated section on the future governance of AI in the military domain. This section calls for the development of a common understanding of AI technology, including its limitations, benefits, risks and consequences. It also emphasizes the importance of open and inclusive discussions among relevant stakeholders, including the private sector, and encourages the identification of complementary synergies among existing efforts on military applications of AI.

Taken together, the REAIM outcome documents represent an incremental and evolving approach to addressing AI in the military domain.

---

105 REAIM, "Pathways to Action", paragraphs 23(d), 23(a), 23(c), respectively.
106 REAIM, "Pathways to Action", paragraph 4.
107 REAIM, "Pathways to Action", paragraphs 15, 18.
108 REAIM, "Call to Action", paragraph 7.18; REAIM, "Blueprint for Action", paragraph 6.8; REAIM, "Call to Action", paragraph 6.10.
109 REAIM, "Call to Action", paragraphs 6.4, 6.8, 6.13, 6.14, 7.18; REAIM, "Blueprint for Action", paragraph 6.9(f); REAIM, "Pathways to Action", paragraph 24.
110 REAIM, "Call to Action", paragraph 6.15.
111 REAIM, "Blueprint for Action", paragraph 5.5.
112 REAIM, "Blueprint for Action", paragraph 5.5.
113 REAIM, "Blueprint for Action", paragraph 5.6.
114 REAIM, "Call to Action", paragraphs 6.9, 6.11, 6.16, 7.21–24.

Political Declaration on Responsible Military Use of AI and Autonomy

The Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy was introduced by the United States during the first REAIM Summit in February 2023 and launched at the United Nations in November 2023. As of 27 November 2024, the declaration had received endorsement from 58 States.[115]

The declaration comprises 10 measures for endorsing States to implement categorized into three groups:

1. Measures to ensure oversight through the establishment and implementation of national policies, procedures and strategies
2. Measures to ensure accountability and responsibility, for instance of personnel involved in or overseeing the development, deployment and use of military AI capabilities
3. Assurance measures to reduce risks and build confidence in the reliability of AI capabilities

Accordingly, three working groups on oversight, accountability and assurance were established, each working group led by two endorsing State. Elements around international law cut across all three working groups.

The Political Declaration on Responsible Military Use of AI and Autonomy offers several elements that may inform the development of future multilateral norms in the military AI domain. The structure and substantive measures of the declaration can provide insights into areas where greater international convergence could be pursued.

Specifically in terms of norms development, the declaration calls for appropriate safeguards to, among other things, "mitigate risks of failures in military AI"; ensure rigorous testing and assurance of military AI capabilities; provide training to the personnel who use or approve the use of military AI capabilities; develop transparent and auditable methodologies and procedures; to actively take steps to minimize unintended bias in military AI capabilities; ensure oversight by senior officials of military AI capabilities with high-consequence applications; and conduct legal reviews on military AI applications.

Paris Declaration on Maintaining Human Control in AI-enabled Weapon Systems

France hosted a two-day AI Action Summit in February 2025 focusing on the future of AI. The summit resulted in a declaration on maintaining human control in AI-enabled weapon systems, among others.[116] The Paris Declaration was endorsed by 27 States.

The Paris Declaration provides specific language around ensuring human responsibility and accountability for the development, deployment and use of military AI applications. The endorsing States agreed to implement "appropriate safeguards relating to human judgment and control over the use of force".[117]

---

115 "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy", US Department of State, 9 November 2023, https://www.state.gov/bureau-of-arms-control-deterrence-and-stability/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy

116 See Elysee, "AI Action Summit", February 2025, https://www.elysee.fr/en/sommet-pour-l-action-sur-l-ia.

117 Paris Declaration on Maintaining Human Control in AI-enabled Weapon Systems, 11 February 2025, https://www.elysee.fr/emmanuel-macron/2025/02/11/paris-declaration-on-maintaining-human-control-in-ai-enabled-weapon-systems, paragraph 2.

The declaration also indicates that endorsing States "will not authorise the decision of life and death to be made by an autonomous weapon system operating completely outside human control and a responsible chain of command".[118] The declaration also acknowledges that international law, including IHL, applies to systems enabled by AI.

Analysis of existing State-led initiatives for norm development

While the scope, ambition and level of participation differ, there are at least seven commonalities across the three State-led initiatives.

First, all three initiatives explicitly affirm that international law, including IHL, applies to military uses of AI. They emphasize that AI-enabled military capabilities must be used in accordance with international law. In this regard, legal reviews of systems that are enabled by AI could help States demonstrate transparency throughout the life cycle of military applications of AI.

Second, and linked to obligations under international law, these initiatives highlight the importance of reducing harm to civilians and civilian objects, calling for the establishment of safeguards to mitigate risks. Several measures, including auditing, evaluations, bias mitigation and risk assessment are addressed by all three initiatives.

Third, a shared core principle that is emerging from these initiatives is the maintain control, human judgment and oversight over the use of force. Following the consensus agreement within the CCW that control and judgment with regard to LAWS are needed to uphold compliance with applicable international humanitarian law, including the principles and requirements of distinction, proportionality and precautions in attack, the importance of preserving control and human judgment in decisions involving the use of force also enjoys broad support across wider multilateral discussions and initiatives within the United Nations.

Fourth, and linked to control, human judgment and oversight these initiatives share similar views on human responsibility and accountability for decisions involving AI-enabled systems.

Fifth, there is broad convergence on the need to develop national policies and legal frameworks to provide adequate oversight throughout the life cycle of military applications of AI.

Sixth, these initiatives recognize the value of training personnel involved in the development, approval and use of military AI systems as part of a broader capacity-building effort at the national level.

Lastly, these initiatives recognize the need for inclusive and multi-stakeholder engagement.

118 Paris Declaration, paragraph 3.

@unidir
/unidir
/un_disarmresearch
/unidirgeneva
/unidir

Palais des Nations
1211 Geneva, Switzerland

UNIDIR.ORG