POLICY BRIEF

Governance of Artificial Intelligence in the Military Domain: A Multi-stakeholder Perspective on Priority Areas

Yasmin Afina • Giacomo Persi Paoli

Acknowledgements

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. This policy brief was prepared by the Security and Technology Programme, which is funded by the governments of Czechia, Germany, Italy, the Netherlands, Norway, the Republic of Korea, Switzerland and the United Kingdom, and by Microsoft.

This policy brief has been prepared to support the work leading up to the 2024 Summit on Responsible AI in the Military Domain (REAIM), due to be held in Seoul, Republic of Korea, on 9–10 September 2024, as well as the UNIDIR's newly launched Roundtable on AI, Security and Ethics (RAISE), both supported by the Republic of Korea and Microsoft.

About UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institute 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 to develop the practical, innovative ideas needed to find solutions to critical security problems.

Note

About the Security and Technology Programme

Contemporary developments in science and technology present new opportunities as well as challenges to international security and disarmament. UNIDIR's Security and Technology Programme seeks to build knowledge and awareness on the international security implications and risks of specific technological innovations and convenes stakeholders to explore ideas and develop new thinking on ways to address them.

About the authors

This report was produced by UNIDIR's Security and Technology Programme.
It was drafted by Yasmin Afina and Giacomo Persi Paoli.

Abreviations

AGI	Artificial general intelligence
AI	Artificial Intelligence
CCW	(Convention on) Certain Conventional Weapons
GGE	Group of Governmental Experts
LAWS	Lethal autonomous weapons systems
IHL	International humanitarian law
IHRL	International human rights law
RAISE	Roundtable for AI, Security and Ethics

Contents

Introduction

As progress in the field of artificial intelligence (AI) proceeds at breakneck speed, the transformative potential of these technologies is bringing to the fore profound implications for national and international security. Consequently, policymakers and regulators worldwide are increasingly recognizing the urgent need for shared understandings that transcend borders and individual interests, particularly in the case of AI's applications in security and defence.

However, the absence of established global collaboration frameworks on AI, security and defence between state and non-state actors poses a significant challenge. This lack of shared governance leads to uncoordinated technological advancement and fragmentation, which has serious consequences for international peace and security, stability, and prosperity. Yet, there is a growing appetite among states to engage with non-state actors. States recognize that this engagement can, at the least, inform governance approaches and solutions and ensure that their development, adoption and implementation are evidence-based. In addition, it can secure awareness and buy-in by industry, civil society organizations, the research, technical and scientific communities, and academia.

This sentiment is shared at the highest levels: in his New Agenda for Peace, the United Nations Secretary-General emphasizes the importance of "ensuring engagement with stakeholders from industry, academia, civil society and other sectors" in the development of "norms, rules and principles around the design, development and use of military applications of artificial intelligence through a multilateral process".[1] There is thus a dire and pressing need for the establishment and promotion of, and support for, an independent, neutral and trusted platform that will enable multi-stakeholder dialogue and incubate governance pathways and solutions for the responsible development, acquisition, deployment, integration and use of AI technologies in the military domain.

1 United Nations, A New Agenda for Peace, Our Common Agenda Policy Brief 9 (New York: United Nations, July 2023), https://dppa.un.org/en/a-new-agenda-for-peace, p. 28.

The Roundtable for AI, Security and Ethics (RAISE)

Recognizing the importance of multi-stakeholder dialogues on AI in security and defence, UNIDIR, in partnership with Microsoft, has launched the Roundtable for AI, Security, and Ethics (RAISE). This is a collaborative multi-year initiative that is intended to foster inclusive, cross-regional and multisectoral engagement. By convening experts from diverse backgrounds, including industry and civil society, RAISE aims to promote open dialogue and cooperation in order to address the complex implications of AI for national and global security.

The inaugural edition of RAISE took place at the Bellagio Center in March 2024. Its objectives were twofold:

(a) Review the current state of applications of AI in security and defence contexts, across sectors and geographies but with a particular focus on the military domain

(b) Identify key priority areas in which to develop specific guidance and policy recommendations on identified issues

This inaugural edition of RAISE focused specifically on the military domain due to the momentum in this particular area. This included the processes and preparatory work in the lead-up to the second Summit on Responsible AI in the Military Domain (REAIM), to be held in Seoul in September 2024, as well as the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, issued by the United States in November 2023 and endorsed by a growing number of other states.[2] The scope of RAISE is, however, intended to gradually cover wider security and defence applications, including AI applications for national security, environmental security and human security.

Facilitated by UNIDIR, the convening of RAISE provided a neutral platform for communities to discuss, to build bridges and trust, and to develop tools to advance ideas and practical actions that contribute to the responsible development, deployment and use of AI in the military domain.

By its end, participants identified six key priority areas for RAISE to advance: building a knowledge base; trust building; the human element in AI uses; data practices; life cycle management; and destabilization.

2 "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy", US Department of State, 9 November 2023, https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/.

Why the multi-stakeholder perspective is important for responsible AI in the military domain

Amid the growing integration of AI into military operations, the international community finds itself at a crossroads, where the development, deployment and use of these technologies must align with legal and ethical imperatives. Yet, a significant portion of the research and development in this space, along with substantial financial investments, lies within the private sector. In addition, academia and civil society organizations play indispensable roles in offering critical perspectives, expertise and evidence on the technological, policy, humanitarian, legal and ethical implications of military AI. A multi-stakeholder approach to informing policy responses and governance frameworks is thus needed and will pave the way for building trust and transcending geopolitical tensions, industrial rivalry and competition.

Multi-stakeholder dialogue offers the opportunity to bridge the gap between communities ranging from the technical via the policymaking to the legal, ethics and humanitarian sectors. Targeted and structured cross-sectoral discussions will subsequently enable a meaningful exchange of knowledge, expertise and experiences with the technology. These discussions will thus improve literacy across communities and actors on the inherently complex and multifaceted implications of AI in the military domain. Primary and secondary beneficiaries of such efforts include states (especially from the "Global Majority"), policymakers, members of the armed forces (i.e., the intended procurers and users of such technologies), the private sector (particularly with regards to compliance requirements), and civil society organizations (i.e., to foster legitimacy, transparency and accountability in the development, deployment and use of these technologies). This will ultimately ensure inclusivity in the development, implementation and operationalization of approaches, norms and principles that may stem from these discussions.

Multi-stakeholder input also lies at the foundation of evidence-based governance approaches and solutions. It ensures that their adoption, implementation and operationalization are feasible, robust, resilient, scalable and sustainable. These approaches and solutions must, in fact, take into account a number of critical factors, including:

- The inherently technical nature of AI
- Military necessity
- The principle of humanity, as well as humanitarian and ethical considerations
- Compliance with applicable laws, including international humanitarian law (IHL) and international human rights law (IHRL)
- The national, regional and international security policy landscape in which these technologies are being developed, deployed and used – including geopolitical divides and (competing) strategic priorities

As such, an established, neutral and trusted multi-stakeholder platform will **clarify the distribution of roles and responsibilities between different actors, organizations and agencies – state and non-state alike** – in implementing and operationalizing responsible behaviour for the development, deployment and use of AI in the military domain.

In addition, the involvement of non-state actors – including industry (from major tech companies and defence contractors to small- and medium-sized enterprises), the research community (academia/universities, think tanks, research institutions and laboratories), as well as civil society organizations and advocacy groups – will **incentivize** their active contribution. For instance, companies will affirm their commitment to responsible innovation, thus ultimately increasing trust among shareholders, policymakers and users. The meaningful inclusion of civil society organizations, advocacy groups and the research community will ensure **public trust and legitimacy**, while fostering transparency and accountability in the responsible development, deployment and use of AI in the military domain.

States will be provided with meaningful input and evidence to consolidate governance solutions and their eventual adoption of norms and principles. Simultaneously, they will leverage the opportunities that AI technologies offer in the security and defence sector, promoting **cross-sectoral innovation, partnerships and collaboration**.

In recognition of the opportunities offered by multi-stakeholder input, UNIDIR, in partnership with Microsoft, inaugurated RAISE to provide a safe, neutral and independent platform for dialogue and trust-building on AI in the military domain. The first edition, in March 2024, laid the foundation for future work that will bring to life recommendations revolving around the six priority themes that the meeting identified (see Figure 1), and which it agreed would serve as a basis for cooperation and collective action that transcends geopolitical rivalry, cross-sectoral divides and competition.

Priority Area 1:
Building a knowledge base

Priority Area 1:
Building a knowledge base

Building a shared and solid knowledge base on AI in the military domain, developed through cross-regional, inter-disciplinary and multi-stakeholder input.

There is widespread recognition of the critical need to promote the responsible development, deployment and use of AI in the military domain. Yet, there is also an understanding that universally agreed-upon definitions are not only unattainable, but that perpetual disagreement can also hinder progress. In its early years, this was an issue with which the Group of Governmental Experts (GGE) on lethal autonomous weapons systems (LAWS) under the Convention on Certain Conventional Weapons (CCW) had to grapple.

Differences in definitions, approaches and terminology at the international level is indeed unsurprising and to be expected. At the national level, states' own policy landscape, strategic priorities, values, ambitions, culture, legal traditions and history will inevitably shape their respective approaches to, and consequent definitions for, critical issues such as AI in the military domain. These divergences subsequently lead to a fragmented policy landscape at the international level, hampering meaningful governance discussions at the multilateral level. This issue has been recognized, for example, by the China–United States Track II Dialogue on AI and International Security led by Tsinghua University's Center for International Security and Strategy and the Brookings Institution.[3]

A knowledge base will ultimately be made from an ecosystem of efforts as outlined in the recommendations below. By navigating through the different approaches that exist across geographies and across sectors, building such a knowledge base will not only help future governance deliberations, but will ultimately build capacity and trust among stakeholders. It is thus important to stress that the establishment of such a knowledge base is not an end in itself but, rather, a critical means to an end: advancing responsible AI in the military domain.

Recommendations

- Develop a "living" lexicon: UNIDIR will lead and coordinate the development, publication and maintenance of a comprehensive lexicon of definitions and approaches in AI in the military domain.[4] The latter should include an overview of how different states, agencies and instruments approach, define and prioritize foundational terms (e.g., "autonomy", "automation", "control"). This glossary must be maintained to factor in new, emerging and evolving definitions over time.

3 Xiao Qian et al., "The China–U.S. Track II Dialogue on Artificial Intelligence and International Security Interim Report", Center for International Security and Strategy, Tsinghua University, 6 April 2024, https://ciss.tsinghua.edu.cn/info/CISSReports/7041.

4 UNIDIR has an established track record of developing resources in the same vein. See, for example, Almudena Azcárate Ortega and Victoria Samson (eds.), A Lexicon for Outer Space Security (Geneva: UNIDIR, 2023), https://doi.org/10.37559/WMD/23/Space/05.

- Develop a “living” taxonomy of risks, harms and concerns: In addition to the lexicon, UNIDIR will also develop a taxonomy of risks, harms and concerns from across communities and regions. The different approaches, boundaries and prioritization across states must be taken into account, as well as additions and changes over time.
- Identify convergence areas and, conversely, divergences for governance solutions: Building on the lexicon and the taxonomy of risks, UNIDIR must then identify areas of shared understandings and convergences and, conversely, areas of disagreement. This effort will unearth promising directions for international agreement and, relatedly, areas that either require further work or that ought to remain at the national level.
- Promote and provide sustainable support for multi-stakeholder, expert discussions: States, industry and other organizations with available resources (e.g., philanthropic foundations) ought to provide support to enable the research and multi-stakeholder meetings that underpin the above-mentioned efforts, ultimately ensuring and maintaining their relevance, accuracy and timeliness. The mobilization of seed funding will be necessary in the short-term (i.e., in 2024–2025), while the provision of long-term support and solutions for its sustainability will also be needed.

Priority Area 2:
Trust Building

Priority Area 2:
Trust Building

Building trust in the technology and in others

Trust is a critical element that lies at the foundation of governance solutions and approaches to promoting responsible AI in the military domain. Understanding and addressing specific concerns to build trust are of particular importance: this will involve the consideration, unpacking and analysis of various layers of trust; incentivizing responsible behaviour; and focusing on concrete use-cases to identify and address sources of distrust.

Three types of trust ought to be considered:

1. Trust in and among states: At the state-to-state level, trust must be established to pave the way for the development of governance approaches and solutions to promote responsible AI in the military domain. In addition, trust in governmental organs is also critical to ensure adequate State stewardship in the context of the high stakes stemming from the development, deployment and use of AI in the military domain. To this end, depoliticization, the exchange of practices, and the exploration and implementation of confidence-building measures will be key. Shared understandings of ethics and trust in different cultures and language contexts constitute another critical factor in establishing trust among states.
2. Trust in the technology: Several layers of trust surrounding the system ought to be unpacked. In addition to trust in the system's consistent reliability, predictability and performance, trust that these will continue in the military and operations contexts in which the technology is designed to be deployed will be critical (i.e., trust in the task). "Success stories" of AI use in conflict must also be highlighted to steer innovation towards the opportunities that these technologies have to offer (e.g., monitoring unexploded munitions, video verification of munition deployment).
3. Trust in the people: In addition to trust in the user/operator, trust must be established with regards to the commander's intent in any decision to deploy and use specific AI-enabled capabilities in specific contexts. Trust must also be established in those involved in the development, testing and auditing (post-deployment) of AI technologies in the military domain, and the extent to which due diligence has been conducted to minimize risks of negligence and risks stemming from their deployment and use. This includes trust in the contractors and vendors that develop these technologies and solutions for their responsible governance (e.g., assurance in suppliers of third-party systems).

Recommendations

- Prioritize incentivization for compliance: States should identify key intervention points and incentives to encourage companies and militaries to adhere to and comply with adopted norms and principles.
- Identify "red lines": In addition to unlawful use-cases, the identification of worries and concerns will be critical. This includes future scenarios where artificial general intelligence (AGI) will be achieved. These "red lines" must, however, be evidence-based and exaggerated fears or dread risks must be prevented: to this end, knowledge, specificity and granularity will be key enablers in mitigating these risks.
- Prioritize regulatory alignment by the public and private sectors: The identification of areas of convergence across international, regional and national positions, policies, standards and best practices will enable regulatory alignment and the consistency and complementarity of governance frameworks developed, adopted and implemented by both the public and the private sectors. Advancing the mutual recognition of regulations between states and sectors will foster compliance and will reduce the associated costs and risks associated with distrust, the growing complexification, and the eventual fragmentation of the wider policy landscape surrounding the development, deployment and use of AI in the military domain.
- Clarify the interpretation of applicable laws: Comprehensive guidance (e.g., a manual) on the different possible interpretations of applicable laws (including IHL and IHRL) must be developed. To this end, states should develop clear, national positions on how to interpret and apply international law in the context of AI applications in the military domain. In IHL, particular consideration must be given to predictability against distinction, proportionality and precautionary obligations. To this end, the development of a compendium of good practices for developing such positions and interpretations may be useful.[5]
- Translate legal and ethical requirements into technical requirements: Building on the previous recommendation, industry and the technical and research communities would greatly benefit from the "translation" of legal and, where applicable, ethical requirements into technical requirements and, when possible, solutions. This set of technical requirements must be compiled into a set of guidelines (e.g., in the form of a manual) reviewed on a regular basis against technological progress, developments in the regulatory landscape, and evolving risks and strategic priorities at the international, regional and national levels.
- Establish mechanisms for compliance monitoring, evaluation, enforcement and verification: These mechanisms can include agreement, internationally, on technical parameters and benchmarks for implementation (e.g., iterative legal reviews; see Box 1) at the national level. Multi-stakeholder input will be key to enable the development of robust, agile and technology-agnostic mechanisms. The establishment and facilitation of information-sharing across regions and passing on best practices from local implementation efforts will be key in fostering and forging trust and building capacity.
- The establishment, promotion and support of cross-sectoral dialogue: States, companies and other organizations with available resources ought to provide support to enable the research and multi-stakeholder meetings that underpin the above-mentioned recommendations. Cross-

5 In the same vein, UNIDIR has recently published such a compendium for cyber. See UNIDIR Security and Technology Programme, Developing a National Position on the Interpretation of International Law and State Use of ICT, A Compendium of Good Practices (Geneva: UNIDIR, 2024), https://unidir.org/publication/a-compendium-of-good-practices-developing-a-national-position-on-the-interpretation-of-international-law-and-state-use-of-ict/.

sectoral dialogue and the conduct of trust-building efforts (e.g., joint exercises) will pave the way for greater compliance with ethical and legal norms in the light of increasing development, adoption, integration, deployment and use of AI technologies in the military domain. These may include the conduct of tabletop exercises involving members of the armed forces, state representatives, the private sector and civil society; as well as the organization of cross-sectoral dialogue and regular reviews of the policy, legal and ethical landscape. RAISE can play a central role in enabling and facilitating the conduct of such efforts and initiatives, ultimately aiming at complementing and consolidating multilateral, regional and national efforts for the governance of AI in the military domain through an independent, neutral and expert platform.

Iterative legal reviews

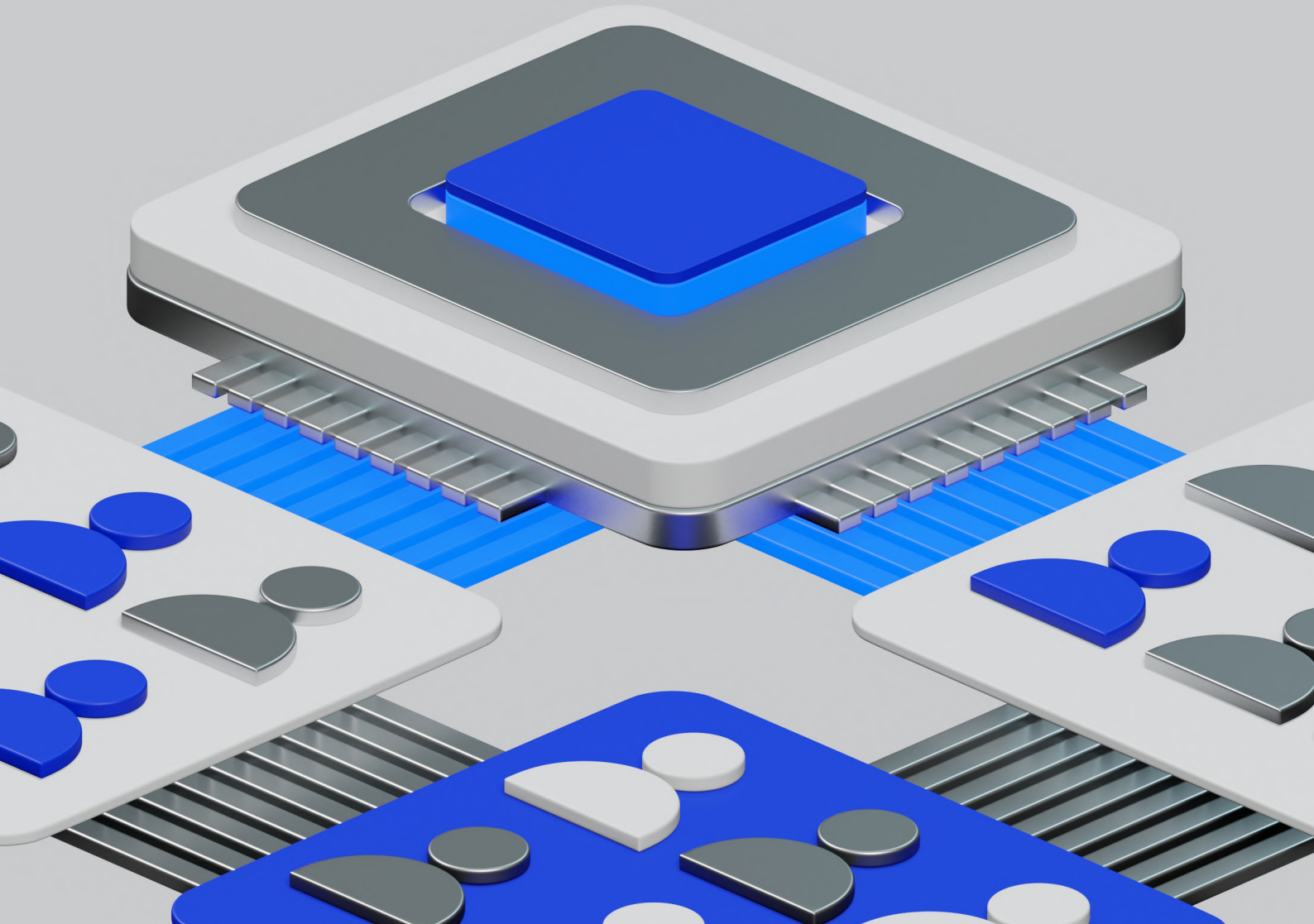
Under international humanitarian law, the choice of means and methods of warfare by parties to an armed conflict is not unlimited. Article 36 of Additional Protocol I of 1977 to the 1949 Geneva Conventions requires state parties to carry out legal reviews in the "study, development, acquisition or adoption of a new weapon, means or method of warfare" to determine "whether its employment would, in some or all circumstances", be inconsistent with applicable laws.[6]

This obligation also applies to AI-enabled weapons and means and methods of warfare. However, the ever-evolving nature of AI means that the way it acts can change. Thus, regular (i.e., iterative) conduct of such legal reviews may be necessary to ensure compliance with applicable laws throughout the technology's life cycle, consistent with states' obligation to respect and ensure respect for the Geneva Conventions.[7]

6 Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977, https://ihl-databases.icrc.org/assets/treaties/470-AP-I-EN.pdf, Article 36.

7 Geneva Conventions I–IV, 12 August 1949, https://www.icrc.org/en/law-and-policy/geneva-conventions-and-their-commentaries#text940076, Common Article 1.

Priority Area 3:
The Human Element



Priority Area 3:
The Human Element

Unpacking the human element in the development, testing, deployment and use of AI systems in the military domain

The human element has featured prominently in most, if not all reflections on international governance approaches to the development, testing, deployment and use of AI in the military domain. For example, within the CCW regime, the term "meaningful human control" quickly gained prominence from the early days of the GGE on LAWS. Beyond meaningful human control and beyond the CCW framework, a number of other terms have mushroomed over the years, including "human in/on/out of the loop" as well as "human oversight". Yet, there is no consensus on the definition and boundaries of these terms. While there is the understanding that a universally accepted approach may not be within reach, one common thread emerges: there is a need for a better understanding of the human element and, more generally, of humanity across a technology's life cycle, as well as its implications for issues such as trust, the performance and reliability of the system, and accountability.

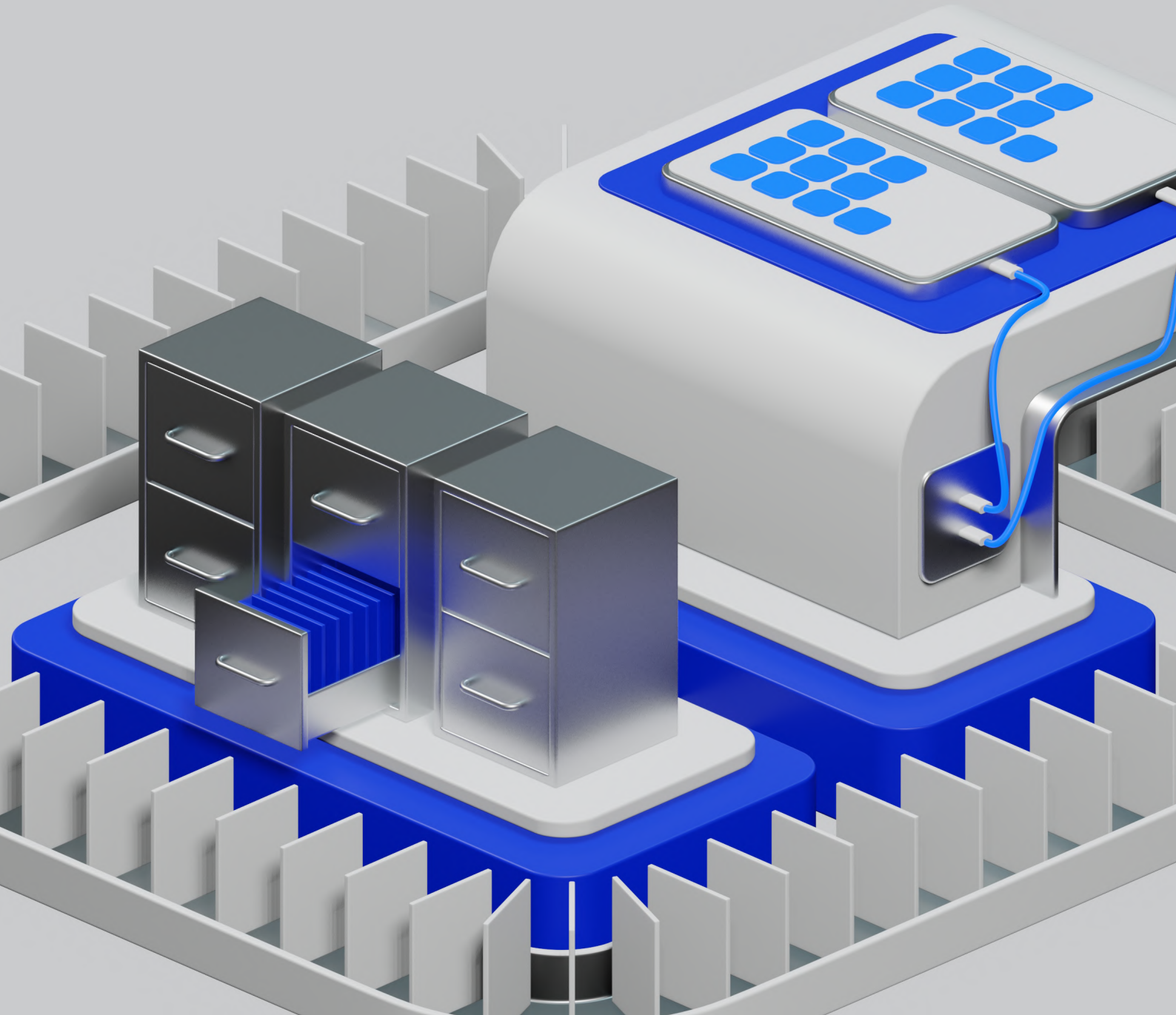
There are many ways through which the human element may be considered relevant across each stage of a technology's life cycle. In the development stages, for instance, the human element can refer to the design of human-centric interfaces and parameters within the technology, thus ensuring the position and role of human operators and their ability to oversee, use or control "by design". The increased automation of certain stages in the development of AI-enabled systems will also need to be examined against its implications, both positive (e.g., to reduce risks of harmful biases) and negative (e.g., reduced oversight and transparency). The human element can also refer to the technical literacy of operators and users and, in addition, awareness of the legal and ethical issues raised by the development, deployment and use of these technologies: training and capacity-building efforts are key enablers of continued and meaningful human–machine interaction across their life cycles. In addition, considerations related to the growing dependency on AI-enabled technologies, as well as the implications of different types of human–machine interactions – including in the context of task delegation in the decision-making chain – merit further attention. Addressing issues of over-hype, over-confidence and over-reliance in the technology will be important, while ensuring trust in the system for its intended use.

These efforts can add tremendous cost to the development, testing, and even procurement and use of these technologies, which not all states may be ready or able to afford. Yet, they are critical to ensuring reliability and, ultimately, accountability: unpacking the human element ensures a clear distribution of roles and responsibilities in line with adjacent factors that may be of influence (e.g., organizational cultures, national policies, regional and national strategic landscape).

Recommendations

- Conduct an in-depth study of the human element: A comprehensive and multidisciplinary study that unpacks the human element in different systems and across their life cycles must be conducted. The study must take into account, among other things, the different ways in which humans have positive and negative influence over the technology's performance and critical points of intervention; it must identify and map best practices and, conversely, deflection practices; must devise technical solutions and benchmarks to ensure legality and permissibility; and must forecast risks, technological progress and subsequent implications (e.g., advent of AGI).
- Integrate the costs of the human element across the technology's life cycle: The development, testing, procurement, acquisition, use and maintenance of AI systems must factor in the resources necessary to consider and ensure the human element.
- Unpack the human element against considerations in international law, including international humanitarian law and international human rights law: The application and interpretation of applicable laws in the development, deployment and use of AI technologies in the military domain can play a critical role in unpacking the human element. As such, the conduct of an extensive study of the human element against legal requirements and considerations will be necessary. This should cover, among other things, the required levels of human oversight, supervision and control over select applications of AI in the military domain (e.g., in the context of military targeting) across the technology's life cycle.
- Establish a global network of scientists and experts for the development of accessible technical solutions: There must be support for the development of technical solutions (e.g., test beds, sandboxes) to ensure the security, controllability, explainability and transparency of AI systems "by design". To this end, a global and multidisciplinary network of scientists and experts should be established, with a clear mandate to develop such technical solutions available to all and at affordable cost; to address harmful biases in underlying data sets; and to identify and address risks of civilian harm. In order to ensure complementarity and avoid duplication of efforts, it must give careful consideration to existing initiatives and efforts, such as those of the Institute of Electrical and Electronics Engineers (IEEE), the International Organization for Standardization (ISO) and the World Wide Web Consortium (W3C). The development, establishment and governance of such a network must be done thoughtfully, and its composition and maintenance must factor in equality, diversity and inclusion considerations. The mobilization of such a network will require considerable time and resources; RAISE can play a critical role in its incubation and establishment through its diverse, cross-regional, cross-sectoral, independent and impartial nature.
- Invest in literacy and capacity-building: Human and financial resources must be dedicated to facilitating the training of operators and military commanders, and ultimately maintaining accountability and responsibility.

Priority Area 4:
Data Practices



Priority Area 4:
Data Practices

Understanding and unpacking data practices for responsible AI in the military domain

A thorough understanding of the layers and issues pertaining to data, their subsequent implications, and how data practices can be leveraged to promote a responsible military AI ecosystem is much needed. Adjacent considerations include the use of data across different domains; their sources; their purchase and transfer; the curation and hygiene practices surrounding the data; as well as the use and means of use of data. Such understanding will enable the identification of issues and, subsequently, governance and technical solutions.

Perhaps the most prominent issue pertaining to data relates to the existence, perpetuation and proliferation of biases. The contextualization of training data sets is thus critical for the reliability of the systems and their ability to comply with applicable laws: for example, the cultural significance of bearing arms can vary, which may affect assessments of targetability and legality. The development and testing of systems relying on such data for target identification must take different forms of biases into account and must address the potential issues that may arise; and users must be aware of these limitations. Other concerns include the explainability of systems and adjacent issues pertaining to the black box nature of systems, as well as the availability of auditing solutions.

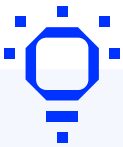
Left unaddressed, data issues may jeopardize compliance efforts: an in-depth study of the subsequent policy, legal and ethical implications is much needed. Such analysis must consider, among others things, risks and unintended consequences; IHL and IHRL; other governance frameworks (e.g., the European Union's AI Act and General Data Protection Regulation); as well as the role and responsibility of the private sector and wider research community (e.g., in the sales of off-the-shelf capabilities trained on untailored data sets).

Recommendations

- Conduct an in-depth study of data practices surrounding military AI: Multidisciplinary and cross-regional research on existing data practices for the development, deployment and use of AI in the military domain is much needed. Such research will enable the development of governance approaches and technical solutions for their operationalization, ultimately enabling responsible AI through data practices.

- Evaluate public–private partnerships and prioritize incentivization: The development, implementation and operationalization of governance solutions around data practices and, more generally, AI in the military domain will require a common effort by public and private entities. This will require a clear distribution of roles and responsibilities, translated through guidelines across the technology's life cycle, from their procurement and acquisition to their use. Incentivization for industry and the wider research community to collaborate and implement requirements and solutions will be key.
- Establish a platform for multi-stakeholder coordination dedicated to data governance in the context of military applications of AI technologies: The establishment of such a platform will enable states and non-state actors to align with and complement efforts to implement governance solutions. This can be achieved, for example, through the sharing of best practices (e.g., for human-centred verification techniques of raw data), ensuring interoperability between systems, as well as by creating mechanisms to validate the quality of data (e.g., sandboxes).
- Establish verifiable standards: The development of standards to complement existing governance frameworks will be critical to address the risks stemming from data practices. Standards and the establishment of benchmarks will play a particularly important role in fostering transparency (e.g., on data transactions, sources and collection, processing, storage, etc.). Implementation efforts must also be established, for example, through data structures for standardization and verification methods.
- Establish iterative legal reviews and human rights impact assessments: Beyond a review of the implications that data practices may have for compliance with IHL, and then work to fit data into the requirements of IHL, it will also be critical to conduct human rights impact assessments on a regular basis in the light of IHRL's applicability both in conflict and in peacetime. These implications include those on the right to privacy, as well as the chilling impact on the freedom of expression and association. Legal assessments must be conducted in an iterative manner (i.e., one-off reviews will not be sufficient; see Box 1) in the light of the ever-learning and, subsequently, ever-evolving nature of AI systems, particularly those that rely on reinforcement-learning techniques (see Box 2).
- Consider complementary technologies for data governance: In the light of the growing body of research on data governance, an exploration and in-depth study of complementary technologies that should be considered for use in the development and verification of AI for military applications. Such technologies include privacy-enhancing technologies that can be used for the training and auditing of models.[8]

8 Elena Himmelsbach et al., "PETs in Practice", Open Data Institute, 22 January 2024, https://theodi.org/insights/explainers/pets-in-practice/.



Reinforcement learning

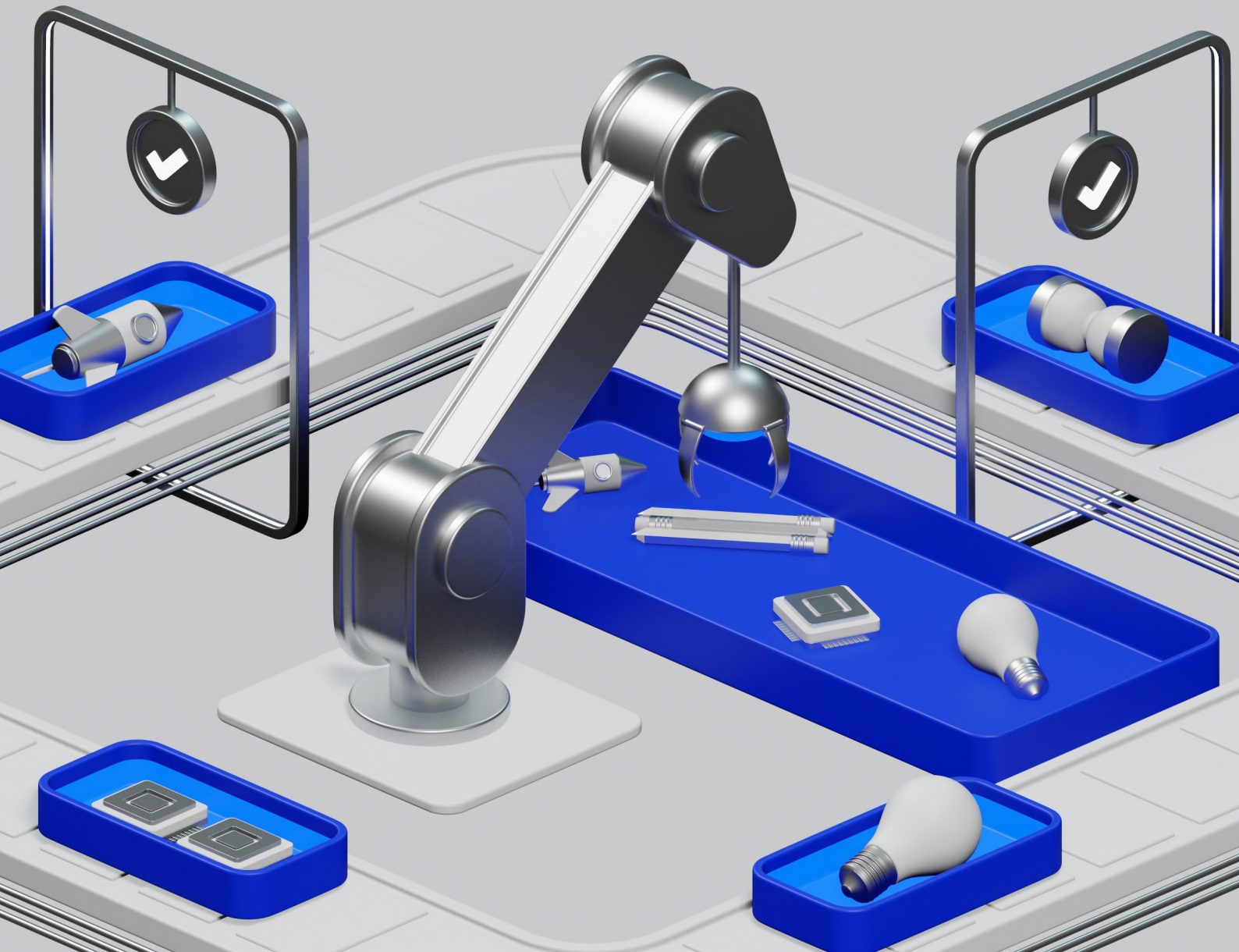
Reinforcement learning corresponds to algorithms that operate on the basis of rewards. Within that context, the algorithm will optimize its performance and determine the ideal behaviour in order to reach the offered reward.[1]

Google DeepMind's AlphaGo is an example of a programme trained on reinforcement-learning techniques: in October 2015 it became the first computer programme to defeat a human professional player in the game of Go; through self-play games, AlphaGo developed its own strategies and optimized the combination and sequences of moves in order to increase its winning rate.[2]

1 Dhairya Parikh, "Learning Paradigms in Machine Learning", Medium, 7 July 2018, https://medium.com/datadriveninvestor/learning-paradigms-in-machine-learning-146ebf8b5943.

2 David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search", Nature 529 (2016): 484–489, https://doi.org/10.1038/nature16961.

Priority Area 5:
Life Cycle Management



Priority Area 5:
Life Cycle Management

Understanding the life cycle implications of AI systems (including the end-of-life) to promote responsible AI in the military domain

Governance approaches centred around the management of technologies across their life cycles to promote responsible AI in the military domain hold much promise. Understanding the different actors, actions and possible points of intervention at each stage of the technology's life cycle would indeed address the whole picture of risks and concerns, most of which revolve around accountability, responsibility and compliance. In its development and testing stages, for example, the adoption of measures to factor in ethical and legal considerations "by design" will be critical to foster compliance. While these may add costs and time prior to the commercialization of products, irresponsible behaviour can ultimately deter clients (i.e., states) in the light of risks of non-compliance and the inability to meet their legal obligations and responsibilities.

In addition, the development and subsequent procurement of an AI technology must also consider the issue of integration and dependency. Once adopted, prospective users (i.e., the military) must be able to integrate this new capability into their wider ecosystem of technologies and assets, including legacy systems. They must ensure interoperability without jeopardizing security against attacks and unintended consequences – for which redundancy measures will be critical especially for systems too critical to fail or too costly to replace. This issue must be considered from the earliest stages of the technology's life cycle, which may prove to be more complex for off-the-shelf capabilities and open-source AI.

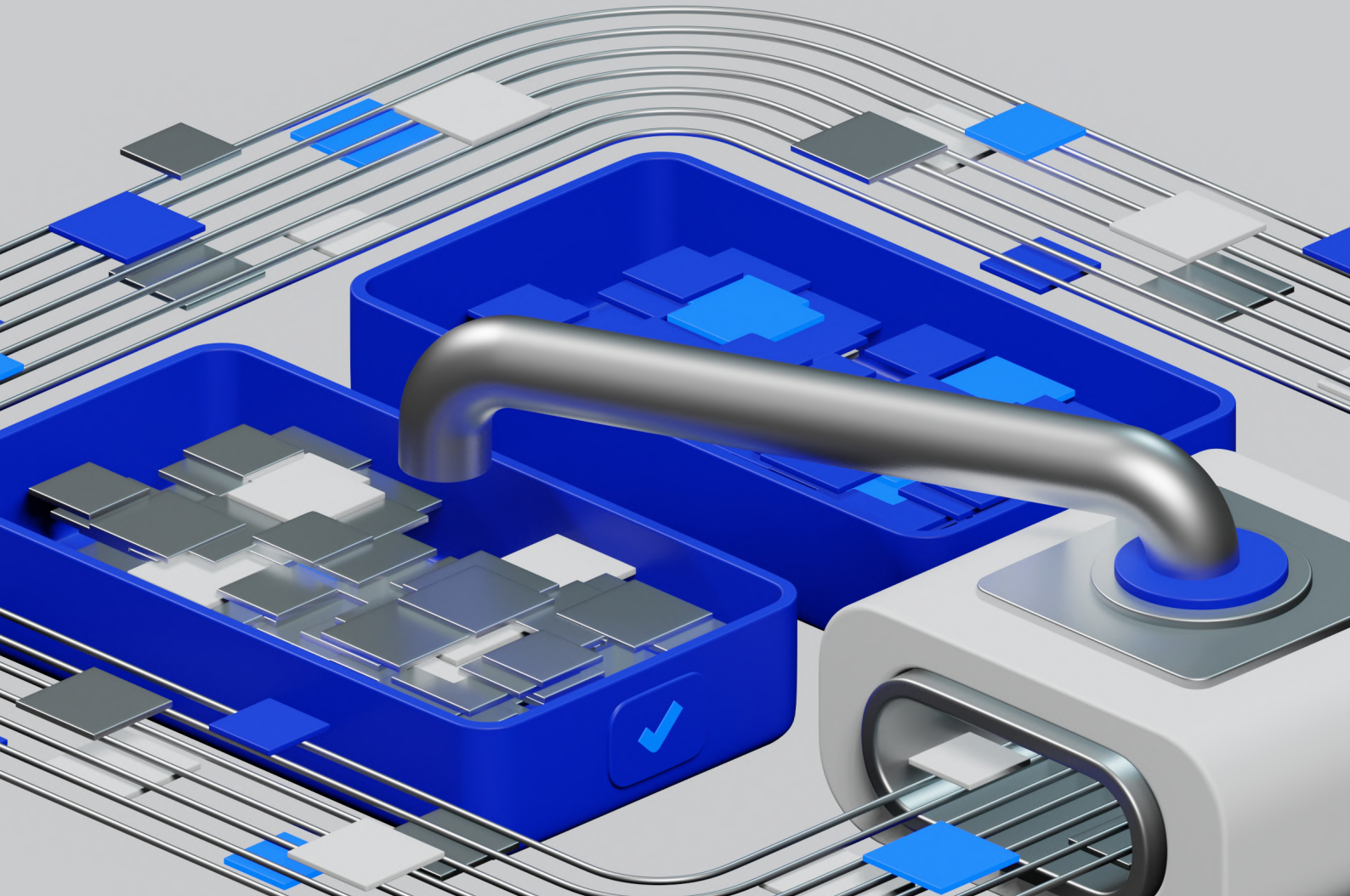
Particular attention must be dedicated to the end of a technology's life cycle, an area that remains much underexplored yet will be critical to address. In the absence of a framework for the governance of a system's final stages, several risks can arise. For instance, in the absence of guardrails, certain actors may resort to irresponsible technology transfer as a disposal strategy that might exacerbate non-proliferation risks. In addition, such practices would jeopardize efforts to ensure responsibility and reliability from the subsequent procurement and use of AI in the military domain, as dated systems may lead to unintended consequences due to its historical usage. Thus, transparency about that past use must be ensured for procurers and users.

How these considerations are to be translated into governance frameworks and operationalization solutions remains to be seen. In the development of such approaches, states, industry, the research community and civil society must take into account ethics and existing legal frameworks, including IHL, IHRL and the United Nations Charter.

Recommendations

- Adopt clear policies and strategies: States, industry and the research community must adopt clear policies and strategies on life cycle management. An iterative, evidence-based review process with multi-stakeholder and multidisciplinary input will be needed to ensure that these policies and strategies remain up to date. The publication of such policies and the exchange of good practices can play a key role in fostering trust and building capacity across regions and sectors.
- Develop procurement guidelines: Procurement and acquisition have been identified as key points of intervention to promote responsible AI in the military domain. As such, guidelines must be developed to assist states, industry and the research community to identify key criteria for compliance, as well as what can be expected of each of them from development and sale to post-transaction (e.g., after-sales assistance in the context of technology transfer).
- Invest in capacity-building and awareness-raising: Greater efforts must be dedicated to ensuring that states, industry and the research community are aware of policy, legal and ethical requirements and have the appropriate capacity to ensure compliance.
- Provide dedicated attention to the end-of-life: Shared and mutually accepted processes for identifying systems that are reaching the end of their life, and their subsequent management or replacement (e.g., through export control mechanisms), will minimize the risks of unpredictable behaviour, loss of control and malicious exploitation.

Priority Area 6:
Destabilization



Priority Area 6: Destabilization

Understanding the drivers, means and methods of and responses to AI-related destabilization, including those enabled, induced and multiplied by AI systems

The development, integration, deployment and use of AI in military contexts must be looked at in conjunction with the destabilizing effects that these technologies enable, induce or multiply. In fact, on the assumption that AI is not only a force-multiplier but also a threat-multiplier, there are a number of issues that may arise and ought to be considered. These include hybrid warfare, proliferation, the fuelling of arms race and competition dynamics, as well as wider societal disruption that may even go as far as being unintentional. Yet, there is also a lack of consensus as to what the drivers, means and methods of and responses to AI-related destabilization are.

As such, the issue of destabilization requires reflection that expands beyond the military domain. Wider security considerations must be addressed, such as what "stability" means in the first place; what are some of the factors that constitute and contribute to that stability; what resilience and risk-mitigation measures are in place; and in turn what are the opportunities and, conversely, threats that AI-enabled technologies may bring. Destabilization also brings to the fore the convergence of AI with other technologies, domains of warfare and security fields (e.g., cognitive warfare, cyber, misinformation and disinformation), as well as the integration of AI into critical national infrastructures and their resilience. This subsequently adds further layers of complexity to pre-existing security issues, such as attribution, accountability and the governance of dual-use technologies.

There is, therefore, a need to acknowledge that destabilization will expand beyond of the military domain and will require more granularity and nuances that reach the wider security field. Frameworks are already in place to address some of the issues raised by AI-related destabilization, such as espionage, subversion and sabotage; however, the ways in which they apply to AI-enabled information, cyber and autonomous threats remain to be determined. In addition, trust- and confidence-building measures in place between states and non-state actors already address some of the concerns raised by destabilization, including through the development of risks taxonomies.

While inherently complex and multifaceted, an exploration of AI-related destabilization offers opportunities for the development of governance solutions and the identification of common ground. In fact, some of the concerns raised by destabilization (e.g., proliferation risks, misinformation and disinformation) are shared widely across regions and across communities, thus potentially paving the way for consensus or, at least, shared understandings of "nightmare scenarios".

Recommendations

- Conduct an in-depth study of the drivers, means, risks and methods of destabilization and responses to them: Such an analysis will pave the way for the development of concrete solutions for the mitigation and reduction of risk. For instance, attribution concerns can draw lessons from watermarking methods of generative AI; and the identification of deterrent factors can reduce risks from states and non-state actors.
- Consider destabilization risks stemming from technology transfer: The proliferation to malicious actors of AI-enabled technologies with military applications must be considered against risks of destabilization. Yet, the development of safeguard measures must strike the right balance with the need for equitable access to AI technologies: the latter particularly constitutes a driving force behind the growth of open-source AI, to which greater security and safety consideration must be dedicated.
- Develop and adopt norms and principles for responsible state and non-state behaviour: Such norms and principles can take the form of soft law (i.e., non-legally binding instruments such as a code of conduct or ethical commitments) or even hard law (i.e., an international treaty), should there be appetite for such instruments. These must include a clear distribution of roles and responsibilities between states, industry, the research community, international organizations, civil society and other non-state actors; measures to build resilience; incentivization of responsible behaviour and, conversely, direct consequences and implications of irresponsible behaviour; as well as clarity on the application and enforcement of existing rules of international law.
- Establishment of a dedicated, multi-stakeholder and inclusive platform to identify and address destabilization risks: Such a platform would allow for inclusive discussions and dialogue across communities. In addition, it would allow for the establishment of a cross-regional, multidisciplinary steering agency or task forces to keep the international community abreast of evolving risks and threats, as well as possible governance and policy pathways and technical solutions.

Conclusion and the way ahead

Multi-stakeholder perspectives provide key inputs that not only feed into policy discussions through evidence and practical solutions; they ultimately ensure that governance pathways are inclusive and practical and span across sectors and regions. To this end, efforts to enable multi-stakeholder and participatory dialogue ought to be advanced in the light of its role in supporting international processes within the United Nations (e.g., in First Committee discussions) and outside (e.g., at the REAIM Summit), as well as regional and national thinking on AI in security and defence.

UNIDIR has a mission to support states, the United Nations, the disarmament policy community and other stakeholders in identifying and advancing ideas and practical actions that contribute to a sustainable and peaceful world. In the light of its autonomous status within the United Nations, UNIDIR provides a neutral, safe and trusted space for further discussions on consolidating governance efforts in AI, security and defence.

As such, UNIDIR will maintain and intensify its multi-stakeholder engagement across communities through RAISE, with a view to gaining granularity in each of the six priority themes for the governance of AI in the military domain. Through constructive dialogue based on trust, collaboration and openness, UNIDIR will also seek to expand its outreach and increase its impact among groups ranging from policymakers, defence and military personnel, legal and ethical experts, to the scientific, technical and engineering communities. Acknowledging the general-purpose nature of AI technologies, UNIDIR will also seek to shed clarity on the wider security and defence implications that far transcend the military domain through multidisciplinary, robust and evidence-based research.

RAISE

Interested in contributing to or supporting RAISE?

You can submit your additional thoughts, insights and contributions or your reactions on any of the six identified areas of priority, or potential support for RAISE, to **sectec-unidir@un.org**