UNODA Occasional Papers, No. 42, June 2024

Governance of Artificial Intelligence in the Military Domain

Dr. Beyza Unal
Ulysse Richard

United Nations

OFFICE FOR DISARMAMENT AFFAIRS
Occasional Papers, No. 42, June 2024

Governance of Artificial Intelligence in the Military Domain

Dr. Beyza Unal
Ulysse Richard

United Nations

ACKNOWLEDGEMENTS

The authors extend their gratitude to the Ministry of Foreign Affairs of Republic of Korea for their funding and support of this research paper. The authors express appreciation to YOUN Jong Kwon, Director-General for Nonproliferation and Nuclear Affairs, and his team members of the Division of Disarmament and Nonproliferation: PARK Eun-jin, Director, LIM Jong-youb, First Secretary, LEE Hahyung, Second Secretary, YUN Da Som, Third Secretary.

For their contribution and insights during the United Nations–Republic of Korea Conference, we would like to thank Yasmin Afina at the United Nations Institute for Disarmament Research; Netta Goussac, Lexbridge Lawyers; Kerstin Vignard, Johns Hopkins University; Andrew Reddie, University of California; Giacomo Persi Paoli, United Nations Institute for Disarmament Research; Vincent Boulanin, Stockholm International Peace Research Institute; Dongyoun Cho, United Nations Institute for Disarmament Research; and Margarita Konaev, Georgetown University. The active involvement of these experts and many others in our work enriched our knowledge.

Our team members at the United Nations Office for Disarmament Affairs, Tania Banuelos Mejia, Olivia Flasch, Charles Ovink and René Holbach, played key roles in reviewing this research paper. In addition, heartfelt appreciation goes to Aaron Junhoung Yoo at the United Nations Office for Disarmament Affairs for his unwavering support, both in the organization of the United Nations–Republic of Korea Conference and during the publication of this paper. Particular thanks to Diane Barnes for skilful editing of this paper.

ABOUT THE AUTHORS

Dr. Beyza Unal is the Head of the Science, Technology and International Security Unit at UNODA. She specializes in the interaction between emerging technology applications and their impact on international peace and security. Before joining UNODA, Beyza was the Deputy Director of the International Security Programme at Chatham House. She formerly worked in the Strategic Analysis Branch at NATO Allied Command and Transformation, taught International Relations at Old Dominion University, and served as an international election observer during the 2010 Iraqi parliamentary elections. Beyza has a PhD from Old Dominion University. She has published numerous articles on international security matters.

Ulysse Richard is a Consultant in the Science, Technology and International Security Unit at UNODA. Ulysse is finalizing his Master's degree in International Security and International Relations at Sciences Po and Peking University, and is a Fellow at the Institute for AI Policy and Strategy, an Analyst at the Spykman Center for Geopolitical Analysis and an Associate Fellow at the Council on Geostrategy. He is part of the first Youth Leader Fund for a World Without Nuclear Weapons cohort, convened by UNODA and the Government of Japan.

DISCLAIMER

The views expressed in the paper are solely of the authors and do not reflect the positions of the UN Office for Disarmament Affairs or Republic of Korea.

GUIDE TO THE USER

The United Nations Office for Disarmament Affairs (UNODA) publishes the UNODA Occasional Papers series to feature papers that deal with topical issues in the field of arms limitation, disarmament and international security, as well as statements made at meetings, symposiums, seminars, workshops or lectures. They are intended primarily for those concerned with these matters in Government, civil society and the academic community.

Symbols of United Nations documents are composed of capital letters combined with figures. These documents are available in the official languages of the United Nations at https://documents.un.org. Specific disarmament-related documents can be accessed through the disarmament reference collection at https://library.unoda.org.

Available in electronic format at https://www.un.org/disarmament/publications/occasionalpapers.

UNITED NATIONS PUBLICATION
405 East 42nd Street, S-09FW001
New York, NY 10017 USA
Email: publications@un.org
Website: shop.un.org

Sales No. E.24.IX.4
ISBN 9789210031745
eISBN 9789211065237
ISSN 0251-9518
eISSN 2412-1258

Printed at the United Nations, New York

Contents

Executive Summary

- Artificial intelligence (AI) that is safe, secure, and trustworthy can help advance the Sustainable Development Goals, improve gender equality, fight illness, guide vaccine design and development, monitor crop moisture and soil composition in agriculture, and provide early warning of climate risks, among other uses.
- AI systems could be used for tasks that are tedious or repetitive or require more endurance, speed, reliability or precision than a human operator can offer. The unique capacities of such systems can make them attractive not only to armed forces but also to non-State armed groups, although such groups may accept substantially lower thresholds for accuracy and reliability.
- AI has transformative potential, for instance, in military logistics and peacekeeping missions, as well as for early warning of threats of violence to local populations. However, as with any technology, developments in artificial intelligence are not free from risks. While some risks are known, there is an elevated level of uncertainty about future risks. This calls for taking precautionary measures to prevent and minimize such risks and putting in place the necessary guardrails.
- Some of the known risks of artificial intelligence include unintended bias and, as a result, discrimination and representational risks, unequal access and an increased digital divide, environmental and social harms, misinformation and disinformation risks, cybersecurity risks, risks to biosecurity, information hazards, misuse, and malicious uses of technology.
- The potential integration of artificial intelligence in the military domain also raises many questions around ethics, legal frameworks, humanitarian concerns, human rights considerations, accountability, safety and security.
- While armed forces may employ AI-enabled systems, for instance as decision-support tools, the duty to comply with international law, including international humanitarian law and international human rights law, rests with humans.
- The inherent uncertainty around the disruptive and destructive potential of AI and the increased risks that it may pose to civilians and civilian objects in armed conflict necessitate further consideration. In this regard, an important commitment from decision-makers would be to leverage AI technology for the protection of civilians, both in peacetime and in times of conflict, as well as to ensure the technology is employed in a way that does not excessively or unnecessarily increase such risks or is used in a way that violates international law.

- Keeping the protection of civilians and civilian objects at the core of technology development would allow decision-makers to work with the AI community to identify both AI safety and security risks.
- Responsibility for taking such precautions does not only rest with States. The private sector, academia, civil society, AI practitioners, designers and standardization bodies should equally share responsibility throughout the life cycle of AI systems. Effective multilateral responses can only arise from the coordinated interaction of these communities.
- The private sector should be guided by corporate accountability and due diligence obligations, along with their obligations under international law, to assess and mitigate any adverse human rights impacts that the design and development of an AI application may cause. States can impose contractual terms on defence companies and the private sector to conduct legal reviews of their products, which would complement States' obligation to conduct legal reviews under article 36 of Additional Protocol I to the 1949 Geneva Conventions.
- It is equally important to engage early-career AI researchers and practitioners in discussions on the potential misuse of civilian AI technologies for military purposes.
- The large number of initiatives on the use of AI in the military domain currently taking place outside of the United Nations underscores the high interest in the topic of AI and its implications for international peace and security.
- Discussions around AI governance in the military domain could include good practices for the design, development and use of AI systems in military contexts; capacity-building initiatives for the Global South; transparency and confidence-building measures; and further development of norms for the responsible military use of AI.
- Relevant but distinct work in various international forums, including under the Convention on Certain Conventional Weapons (CCW) on lethal autonomous weapons systems, can help create further synergies.
- Once tailored for AI, lessons from other disarmament mechanisms — such as confidence-building measures, de-escalation mechanisms and good practices to exchange timely information — could help reduce tensions and mitigate misunderstandings and misperceptions in the context of international peace and security.

Introduction

There is no multilaterally agreed definition of AI. Various definitions in use refer to "machines with the ability to learn, solve problems, make predictions, take decisions and perform tasks that are considered to require human intelligence."[1]

In the past year, developments in artificial intelligence (AI) have made headlines across the globe, with significant milestones reported in the areas of both general and narrow-focused capabilities.[2]

Notable advances in AI, including in its subfields such as machine learning, natural language processing and computer vision techniques, pave the way to the possible integration of AI-enabled systems into different sectors, such as health care, education, law enforcement and defence.

Breakthroughs in AI technology have been possible mainly due to the exponential growth in computing power, the availability of large — albeit non-inclusive — datasets, novel machine-learning techniques, increased private sector interest, and innovation as a result of the convergence and interaction of AI with other technologies.[3] The wide range of processes that can potentially be optimized using AI, including improvements in efficiency, increased autonomy, and analytical capabilities, makes the integration of AI appealing for a diversity of applications, both in the civilian and military domains.

AI that is safe, secure, and trustworthy can help advance the Sustainable Development Goals, contribute to gender equality, improve conflict analysis, fight illness, guide vaccine design and development, support mediation activities, and provide early warning of climate risks, among other uses.

While civilian applications of AI have received the most attention, especially in the wake of the generative AI boom of recent years,[4] many leading experts, from the private sector to academia, have warned about the international peace and security risks that may arise from AI technologies.[5] While some of these risks are known, others would require further study.

1 See United Nations General Assembly, Current developments in science and technology and their potential impact on international security and disarmament efforts, seventy-eighth session, A/78/268.

2 There has been a shift from research and development for AI to execute specific tasks to building AI models that can perform a wide range of applications. While general-purpose AI systems can perform a broad range of tasks, narrow systems focus on a single task. As outlined by Google and Google DeepMind, "LLMs [language learning models] and the current generation of AI models are characterized by more 'general' capabilities than those of more narrow, non-generative systems like image recognition models – which are adept at a more limited set of tasks oriented towards one capability." See, Google and Google DeepMind, Large Language Models, Written Evidence, UK House of Lords Communications and Digital Select Committee Inquiry, September 2023.

3 For a read on new and emerging technologies, see Suleyman Mustafa and Michael Bhaskar, September 2023, The Coming Wave: Technology, Power and the Twenty-First Century's Greatest Dilemma, Crown.

4 For a brief review of what is meant by generative AI, see MIT News, "Explained: Generative AI", 9 November 2023.

5 See for instance, the call for a pause on powerful AI experiments: Future of Life Institute, "Pause Giant AI Experiments: An Open Letter".

There is a growing need to examine AI in the context of international peace and security, including its life cycle in the military domain as well as the potential diversion and misuse of civilian AI technology for malicious purposes by States and non-State actors.

As the speed of developments in AI outpaces intergovernmental processes to establish necessary guardrails, this paper aims to increase awareness and knowledge of responsible AI in the military domain, highlight areas where there is contention at international forums, and provide options and recommendations for multilateral governance.

The speed of developments in AI outpaces intergovernmental processes

As part of this research paper, the United Nations Office for Disarmament Affairs (UNODA) and the Republic of Korea (ROK) held a two-day conference (4–5 December 2023) in Geneva. The conference was held as part of the twenty-second iteration of the United Nations–Republic of Korea Joint Conference on Disarmament and Non-Proliferation Issues. The conference welcomed around 150 participants and featured 39 distinguished speakers and chairs[6] to discuss the governance of artificial intelligence in the military domain. The conference was held under the Chatham House Rule to maintain an inclusive and open dialogue.[7] This paper includes additional desk research and a literature review to provide a comprehensive analysis of the subject.

This paper will first introduce the risks and opportunities of integrating AI in the military domain. Second, it will provide an understanding of both weapons-specific uses and non-weapons-related uses of artificial intelligence. Third, it will introduce considerations around legal implications. Fourth, it will provide lessons and good practices from civilian applications of artificial intelligence, including ethical and technical considerations. Lastly, the report will conclude with policy recommendations.

6 The speakers were selected based on their expertise in the subject matter and represented a wide range of roles across government, civil society, academia, and the private sector.

7 The Rule reads as follows: "When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed." For more on the Rule, see Chatham House Rule.

1 Risks of AI in the Military Domain

The widespread adoption of AI-enabled capabilities by armed forces, along with recent advances in other emerging fields (e.g., autonomy, data, modelling and simulation), will likely change the nature of armed conflict. Military applications of AI may pose both opportunities and risks, depending on how, where, by whom, and for how long AI is used and on the context in which it is used (i.e., whether in peacetime or a time of conflict, and whether in the domains of land, sea, air, cyberspace or outer space).

Opportunities

AI-enabled systems could perform repetitive and/or arduous tasks requiring endurance, speed, reliability or precision exceeding the capabilities and cognition of a human operator.[1]

AI could enhance situational awareness,[2] for instance, for the early detection of the movement of military forces. It could be used as a decision-support tool to enable better and more timely decisions at the strategic, operational, and tactical levels. AI-enabled systems could also be employed to detect cyber threats. Apart from existing uses of AI for cyber defence, there is ongoing research assessing the potential to employ AI not only for detecting malicious activity but also autonomously responding to it.[3]

With the availability of large datasets, AI could also be used in modelling crises or simulating conflicts for the decision-makers to better understand the environment, perception, or actors' calculations both in peacetime and in times of conflict.[4] The usefulness of these modelling and simulation tools is not yet known. Furthermore, the use of synthetic training environments for training armed forces can combine artificial intelligence techniques with virtual and augmented reality.[5] Integrating AI into education and training, such as through computer-assisted or simulation-based instruction, could enable personalized military training and performance assessment, enhancing instructional efficiency, collective development and cost-effectiveness.[6]

1 United Nations, General Assembly, Current developments in science and technology and their potential impact on international security and disarmament efforts: Report of the Secretary-General, A/78/268, 2023.

2 On the benefits of AI in military transformation, see Verbruggen and Maaike, "The Extensive Role of Artificial Intelligence in Military Transformation." The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk. Stockholm International Peace Research Institute, 2020.

3 Lohn A., Knack A., Burke A., Jackson K., June 2023, "Autonomous Cyber Defence", Centre for Emerging Technology Security, Alan Turing Institute.

4 See for instance an example from wargaming and modelling, Davis P., Bracken P., 15 February 2022, "Artificial Intelligence for Wargaming and Modeling", Journal of Defense Modeling and Simulation: Applications, Methodology, Technology (2022).

5 DSEI, June 13, 2023, "What is a Synthetic Training Environment".

6 Fletcher, J. D. "Education and Training Technology in the Military". Science 323, no. 5910 (January 2, 2009): 72–75.

AI-enabled systems can also be used in data processing and analysis in the military domain, for instance in predictive logistics and maintenance. A potential area of use of AI includes predictive analysis on the status of fuel or ammunition.[7] Moreover, AI could improve predictive analytics and assess inventory needs or delivery times in the military supply chain.

Experts indicate that AI-enabled systems could potentially increase battlefield transparency. The mounting of high-resolution sensors aboard satellites, uncrewed systems and ground-based instruments has enlarged the quantity and quality of readily available data gathered for intelligence, surveillance and reconnaissance (ISR) purposes beyond the limits of human processing.

It is a humanitarian imperative and a moral opportunity to leverage AI applications for the protection of civilians in peacetime and in times of armed conflict. AI-enabled technologies can be an effective tool for early warning to assess threats and levels of violence against civilians.[8] In peacekeeping missions, trustworthy AI technology could help analyse data and present conflict dynamics in an operator-intelligible manner.[9] It can also help monitor compliance with ceasefires and peace agreements in post-mission settings by tracking troop movements, weapons deployments or human rights violations.[10] Natural language processing tools can also support humanitarian action.[11] Such tools demonstrated potential for translation between different languages, which could be used in mediation processes or in communicating with locals.[12]

These potential benefits must be weighed against existing technical limitations and challenges with AI, including data quality, bias, privacy, explainability, transparency, robustness and fairness. Relying on datasets or a machine-driven analysis, for instance, may be problematic, especially considering that "peacetime datasets may not be adequate for wartime operations."[13] Moreover, there are questions about the collection of large datasets for military purposes, especially in regard to the inclusivity of such datasets, data collection methods, and whether the end users have granted access to their personal data, among others. The armed forces also rely on both open-source and classified information in designing and using AI models in the military domains, making public scrutiny harder.

Leveraging AI to protect civilians is a humanitarian imperative and a moral opportunity

7 South T., 7 February 2024, "Future Soldier Resupply Could Rely on AI-powered Logistics, Robo-Boats", Army Times.

8 Sarfati, Agathe. "New Technologies and the Protection of Civilians in UN Peace Operations". International Peace Institute, 5 September 2023.

9 Pasligh H., 17 July 2019, The Application of Artificial Intelligence for Peacekeeping, Security Distillery.

10 See Clément S. G., 2022, "Exploring the Use of Technology for Remote Ceasefire Monitoring and Verification", United Nations Institute for Disarmament Research.

11 See, United Nations Peacekeeping, 15 August 2021, Strategy for the Digital Transformation of UN Peacekeeping.

12 For instance, the United Nations is reportedly working with the AI startup Remesh to facilitate dialogue with local populations in conflict zones, such as Libya and Yemen. See Brown, Dalvin. "The United Nations Is Turning to Artificial Intelligence in Search for Peace in War Zones". Washington Post, 16 April 2021.

13 Zhang, Li Ang, Yusuf Ashpari, and Anthony Jacques. "Understanding the Limits of Artificial Intelligence for Warfighters: Volume 3, Predictive Maintenance". RAND Corporation, 3 January 2024.

Figure 1. AI Application Areas and Corresponding Existing and Potential Tasks: A Selective Overview

Applications	Existing & Potential AI-enabled Tasks
Command, control & communications	• Target selection • Launch of an attack / Engagement
Intelligence, Surveillance & Reconnaissance	• Data collection, fusing, analysis, and dissemination for enhanced situational awareness across domains • Decision support • Automated classification and/or assessment of sensors data • Tactical reconnaissance • Real time threat assessment • Detection of hidden/camouflaged targets
Logistics	• Predictive maintenance • Resupply of goods to the battlefield • Supply chain management • Inventory management • Deployment of people and equipment
Autonomy	• Precision guidance of weapons • Swarm technology, aerial surveying • Navigation and path planning • Avoiding obstacles • Search and rescue missions in remote areas • Target identification and object detection • Adaptive learning
Medical Assistance	• Injury diagnosis • Automated triage • Treatment and health monitoring • External expertise guidance
Modelling & Simulation	• Training of personnel for combat/hostile environments • Virtual reality or simulated environments for mission rehearsal

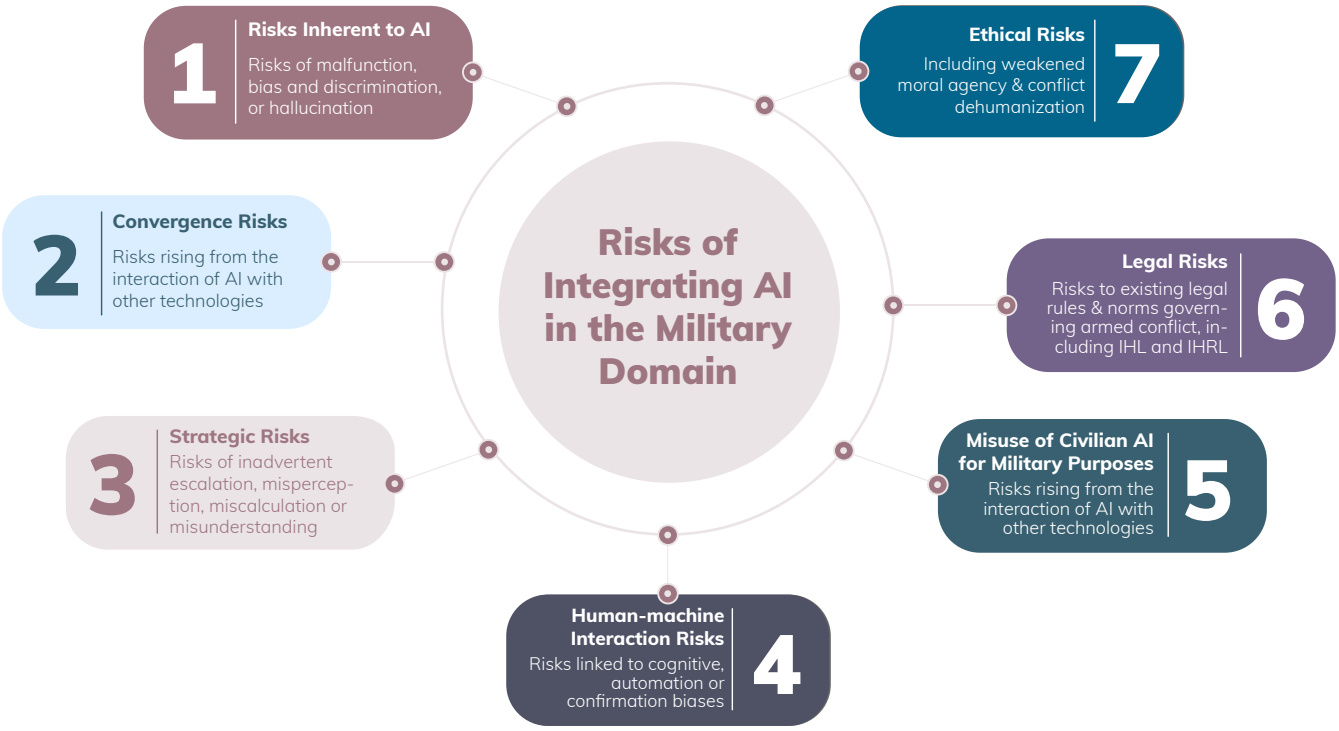
Risks

The use of AI in a high-stakes context, such as in armed conflict or in an environment with civilian presence, requires particular attention not only due to potential risks it poses to international peace and security but also due to considerations under international law. While some risks are foreseeable (i.e., known risks), AI could also exacerbate existing risks and create new ones with unintended consequences. Existing and potential risks can be classified into several categories:

- **Risks inherent to AI technology**, including malfunction, bias, discrimination and representational risks
- **Risks arising from the interaction of AI technology with other technologies**, including in the areas of misinformation and disinformation, cybersecurity, biological threats or information hazards
- **Strategic risks** related to the use of AI in the military domain, including risks of escalation, miscalculation, and misunderstanding
- **Risks posed by how operators and decision-makers interact with the technology** and as a result of human-machine teaming (e.g., the role of human operators in decision-making in light of cognitive and confirmation bias)
- **Risks of misuse of civilian AI technology** by States and non-state actors and malicious uses of civilian AI technology for military purposes
- **Legal risks** arising from existing legal rules and norms that apply to all parties in armed conflict
- **Ethical risks**, for instance, in relation to the deliberate targeting of civilians by an AI system

This section provides a detailed summary of some of these risks.

Figure 2. Risks of Integrating AI in the Military Domain



Risk of Malfunction

Numerous unaddressed safety risks can lead AI systems to malfunction during peacetime and in times of conflict. Risk scenarios include the misalignment between an operator's intent and a system's actual behaviour[14] (***alignment***), a system's lack of resilience to unusual situations or events (***robustness***), an operator's inability to detect unexpected model outcomes or functionalities (***monitoring***), or issues emerging from the broader context in which AI systems are handled (***systemic safety***).[15] Since systems solely rely on statistical associations learned during the training phase, slight changes in the operating environment can affect their performance. As AI and machine learning systems will increasingly be used in high-stakes environments (e.g., on the battlefield or through integration into critical infrastructure), their safety risks can also lead to security concerns.

Moreover, experts indicate that current AI models are brittle – meaning that algorithms cannot generalize or adapt to conditions outside a narrow set of assumptions. Testing, evaluation, validation and verification of these systems for all possibilities is impossible.[16] Brittleness becomes a big problem in safety-critical systems such as critical infrastructure or weapons platforms — errors in such systems could endanger human life.

Military operations that rely on AI-enabled systems may exacerbate existing risks to international peace and security and create new ones. Such operations may result in "unforeseen system failures", have "cascading effects triggering accidental escalation", or lead to inadvertent or deliberate escalation.[17]

Risk of Bias

The notion of AI bias,[18] understood broadly, can encompass three main categories: systemic, human, and statistical and computational bias.[19] Each type of bias can produce harmful outcomes. Training an AI-enabled target identification system with incomplete or non-representative datasets, for instance, could lead to the misidentification of civilians or combatants in armed conflict — thereby posing legal concerns around the principle of distinction under international humanitarian law.

Unintended bias often results in discrimination, and it is often felt by marginalized or minority groups the most. In other words, when an AI-enabled system misclassifies an individual or makes unfair

14 To better conceptualize the challenge of misalignment, consider an example drawn from the self-driving car industry. "While a self-driving car may be programmed to maximize fuel efficiency by taking a longer road to avoid traffic, it might make you late for a meeting. "In this case, the AI system's goal — fuel efficiency — is misaligned with your personal goal of arriving on time." See, Knowledge Zone.

15 Hendrycks, D, Carlini N., Schulman J., and Steinhardt J. "Unsolved Problems in ML Safety." arXiv, June 16, 2022.

16 Cummings, M. L. "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings". AI Magazine 42, no. 1 (2021): 6–15.

17 See Hoffman, W, and Heeu M. K., March 2023, "Reducing the Risks of Artificial Intelligence for Military Decision Advantage". Center for Security and Emerging Technology.

18 Bias refers to deviation from a standard. Despite its negative connotation in the English language, voluntarily introduced moral choices can introduce bias but might ensure adherence to moral norms. For instance, 'an autonomous weapons system might be provided with an ethical regulator that will not allow it to fire at perceived enemy combatants if they are near a UNESCO-protected historical site." See Danks, David, and Alex John London. "Algorithmic Bias in Autonomous Systems". In Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 4691–97. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, 2017.

19 Schwartz, Reva, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence". Gaithersburg, MD: National Institute of Standards and Technology (United States), March 15, 2022.

predictions, for instance, based on age, race or sex, those with the least power frequently bear the brunt.[20, 21] Ensuring AI systems are equitable and just could help to address systemic inequalities.

Unintended bias can be featured in any application that relies on large datasets or algorithms. Data quality, therefore, is an important feature of AI reliability but remains a critical bottleneck. Accurate AI-enabled outputs require large quantities of representative and contextual datasets. Data is an important component of not only training an AI-enabled system but also when testing the system; and the initial training dataset should also be different from the testing dataset.

Bias is also a feature of humans (e.g., human prejudices). Some researchers argue that it is not possible to expect AI to be rational and non-biased when humans, themselves, are biased; however, a distinction holds: Humans can explain their decision-making processes and would be responsible and accountable for their decisions.

Technology is not an antidote for bias. There is a general agreement within the AI community that instead of ignoring unintended bias in data and algorithmic choices, such bias should be acknowledged and controlled for — for instance, through de-biasing methods. While controlling for all types of bias may not be possible, ruling out instances that may cause extreme deviations from the desired outcome is necessary for a dataset to function as intended.[22] Identifying and mitigating unintended bias is central to ensuring the responsible life cycle of AI in the military domain.

Lack of Transparency and Interpretability, or the "Black-box" Problem

The functioning of machine learning algorithms can be opaque to their designer. This issue is often referred to as the "black-box" problem: while inputs and outputs are observable, the steps of the algorithmic process often remain unknown — particularly with deep learning models.[23] AI models are complex systems with many levels of connections and hidden patterns between algorithms. With a high level of complexity comes a lack of transparency, interpretability, and unpredictability. Often, it is unclear how a model produces a proposed output (e.g., target selection, threat assessment).

20 An often-cited example is the misclassification of black people as gorillas by Google Photos in 2015 – this error resulted from the image recognition algorithm used by Google. In 2023, the New York Times tested Google's algorithm and those of its competitors, reporting that the problem persists and that Apple faces a similar problem. See, Grant N. and Hill K., May 22, 2023, "Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's", The New York Times; See also Barr A., 2015, "Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms", The Wall Street Journal.

21 In a 2018 examination of facial-analysis commercial software (IBM, Microsoft, and Face ++), MIT Media Lab's Joy Buolamwini showed a significant disparity in error rate across genders and skin colours (i.e., error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women). See Buolamwini, Joy Adowaa. "Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers". Thesis: S.M., Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences, 2017.

22 One school of thought indicates that de-biasing data often introduces human preferences and fails to address the root cause of the problem. In this regard, some researchers argue that efforts to address bias should focus on end users rather than the training datasets.

23 Pedreschi, Dino, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. "Meaningful Explanations of Black Box AI Decision Systems". Proceedings of the AAAI Conference on Artificial Intelligence 33, no. 01 (17 July 2019): 9780–84.

A significant concern is the absence of a contestability mechanism over the actions and decisions of AI-enabled systems.[24] Contestability rests on the notion that humans should dispute algorithmic decisions. It is, in essence, a design objective that can contribute to the humans' understanding of AI-enabled systems.[25] Contestability, when successfully incorporated, can help keep human agency in the decision-making process.

From a legal standpoint, the lack of insight into how an AI system arrives at a conclusion poses challenges to accountability. Humans must possess sufficient information about critical decisions to ensure their actions comply with international law. Black-box issues could also take the form of unforeseen behaviours; for instance, an AI system may learn unexpectedly not to allow human operators to override its decisions.

With a high level of complexity comes a lack of transparency, interpretability and unpredictability

Strategic Risks

An additional layer of risks arises from the ongoing technological competition between States and private sector actors. A self-imposed sense of urgency to capitalize on AI's strategic advantages can lead to sloppy practices around adherence to ethical and safety standards, potentially resulting in the "use of novel technologies without putting the necessary humanitarian and safety constraints in place."[26]

The diffusion of AI technology to new actors (both States and non-State actors) can lead to the expansion of existing threats (e.g., cyber threats), introduction of new threats, and transformation of the characteristics of these threats.[27]

A key strategic concern in the military domain is the speed at which AI-enabled systems operate, potentially leading to compressed decision-making time frames. The use of AI in critical functions may lead to compressed time frames, which could result in increased tensions, miscommunication and misunderstanding among States. While scientific developments may allow military forces to build systems that can act and react faster than human cognition, this does not mean that decision-makers should "automate" decision-making through machines. In fact, slowing down the processes, by putting necessary guardrails to ensure meaningful human control and/or oversight over the use of AI-enabled systems in high-risk contexts, would protect commanders and decision-makers at large from causing unnecessary and unintended harm to civilians.

Systems capable of processing data at machine speed without verification of their actions by human operators could lead to severe consequences. For instance, AI-enabled systems may end up responding to false-positive incidents or engage civilians and non-hostile targets

24 Contestability refers to features allowing end users to contest algorithmic decisions, fostering active questioning instead of passive acceptance of AI-driven outcomes. See Mulligan, Deirdre K., Daniel Kluttz, and Nitin Kohli. "Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions". SSRN Scholarly Paper. Rochester, NY, 7 July 2019.

25 Ibid, p. 9.

26 Morgan, F. E, Benjamin Boudreaux, Andrew J Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. "Military Applications of Artificial Intelligence," p. 48. Santa Monica: RAND Corporation, 2020.

27 Johnson, James. "Artificial Intelligence & Future Warfare: Implications for International Security". Defense & Security Analysis 35, no. 2 (2019): 147–69; Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation". arXiv, February 20, 2018.

erroneously. In situations involving human-machine teaming, the opacity and speed of AI-enabled decision-making can also pose challenges to ensuring human control over the use of force.

The integration of AI-enabled systems into strategic weapon systems could also increase the prospects for miscalculation and mistake, potentially endangering the protection of civilians in armed conflict. While earlier technologies in history had well-understood operational limits, often supported by physically oriented safeguards or verification, the functioning of AI and machine learning applications often eludes human comprehension.[28]

AI-enabled capabilities may lead to deliberate, inadvertent or accidental escalation in times of conflict.[29] Considering that an AI application can only be trained for a number of scenarios and that there will always be unforeseen cases that fall outside the model's defined parameters, it is likely to expect problems in its actions. Moreover, humans — operators, coders and AI practitioners in general — may not necessarily "know that [a machine learning] application is dealing with a novel scenario."[30]

Risks of Convergence with Other Technologies

While some risks are known, there is an elevated level of uncertainty about future risks. The convergence of AI with other developments in science and technology — such as in biology, quantum, and cyber technologies — is of concern.

Experts indicate that the proliferation of AI-enabled cyber capabilities will likely change the cyber threat environment. For instance, phone frauds are more convincing today because scammers use deepfakes to mimic the voices of loved ones.[31] Large language models (LLMs) can be repurposed to write patches or find coding exploits. And, while the AI expert community comes up with solutions, such as a watermarking technique for images that verifies real content, there are areas that still lack clear solutions (e.g., verification of the authenticity of audio content).

AI models could also be subject to adversarial attacks. Adversarial machine learning techniques may break down the integrity, confidentiality and privacy of datasets without the knowledge of end users.[32] Such attacks could occur on the input, output or training datasets.[33] Moreover, the injection of malicious or misleading data into training datasets could lead to hallucinations or the model to provide spoofed outputs.

28 Lin, H. 2019, "Escalation Risks in an Artificial Intelligence—Infused World." Artificial Intelligence, China, Russia, and the Global Order. Air University Press, p. 150.
29 Ibid.
30 Ibid.
31 Flitter E., Cowley S., 30 August 2023, "Voice Deepfakes are Coming for Your Bank Balance", The New York Times.
32 This is known as data poisoning.
33 For other types of attacks, see Vassilev and Apostol. "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations". Gaithersburg, MD: National Institute of Standards and Technology, 2024.

The potential integration of AI into nuclear weapons systems may further increase the risk by introducing new nodes of vulnerability and escalation pathways.[34] In particular, the use of AI in the pre-delegation of the launch of nuclear weapons could result in catastrophic outcomes.

There is a risk that advances in science and technology reduce the technical barriers to the development of biological weapons. Deadlier pathogens can be made from scratch. In the area of chemical weapons, for instance, researchers found that AI could identify 40,000 potentially lethal molecules in just six hours.[35]

The interaction of neurotechnology[36] — a subfield of biotechnology — with AI also requires increased attention. Both the private and defence sectors[37] are investing in civilian and military applications of human enhancement and degradation technologies that utilize human-machine interfaces. While such technologies can be used for good to improve health and daily life, they could also be used for malicious purposes and in armed conflict. These technologies could also have an "impact on the ability of the soldier to follow the law of armed conflict.[38]

With the introduction of AI and increased autonomy in decision-making, human-machine interaction (also referred to as human-machine teaming) is also becoming complex. There are psychosocial[39] and sociotechnical[40] impacts of human-machine teaming. Some of the challenges that arise from those impacts include "illusion of control, heuristic shortcuts in decision-making [and] automation bias."[41]

The convergence and interaction of AI with other technologies can result in unexpected outcomes, including cascading impacts across critical infrastructure or in high-stakes environments.

34 On AI and nuclear weapons, See Boulanin, Vincent. "AI & Global Governance: AI and Nuclear Weapons - Promise and Perils of AI for Nuclear Stability". United Nations University - Center for Policy Research, 2018.

35 Calma J., 17 March 17 2022, "AI suggested 40,000 new possible chemical weapons in just six hours", Verge.

36 Research around enhancement includes different techniques, including cognitive enhancement that improves one's intelligence or memory; a physical enhancement that enhances performance or endurance; emotional enhancement that improves or controls one's mood; moral enhancement that corrects behavior; cosmetic enhancement that alters bodily appearance; and longevity enhancement that slows down aging. See Erden Yasemin and Brey Philip, "Sienna D5.3: Methods for promoting ethics for human enhancement".

37 Some military applications include light and durable exoskeletons, prostheses for endurance and strength, new sensors, cognitive enhancements, brain-machine interfaces connected to command, control and communication systems, and medical countermeasures for stress resistance of soldiers or treatment of post-traumatic stress disorder.

38 See Henschke A., 3 July 2017, ICRC Blog, "'Supersoldiers': Ethical concerns in human enhancement technologies".

39 "Psychosocial" in this context refers to the interrelation between individual cognitive or mental factors and social factors.

40 "Sociotechnical" in this context refers to the interrelation between social factors and technological developments.

41 Johnson, James. "The AI Commander Problem: Ethical, Political, and Psychological Dilemmas of Human-Machine Interactions in AI-Enabled Warfare". Journal of Military Ethics 21, no. 3–4 (October 2, 2022): 246–71.

Figure 3. Risks Arising from the Interaction of AI with Other Technologies

2 Integration of AI in Weapons Systems

AI will likely be integrated into the military domain at the strategic, operational, and tactical levels.[1, 2] Depending on the task at hand, military applications will leverage the core capabilities of AI technology, including but not limited to the ability to process, interpret and classify visual information (i.e., computer vision) to enable computers to understand, interpret and generate text or voice data (i.e., natural language processing), to analyse large volumes of data to predict future outcomes and craft tailored strategies (i.e., problem-solving and planning).[3]

The use of AI in military systems can be organized under two categories: (a) weapon-specific uses of AI in military systems (e.g., for targeting and attack purposes), and (b) non-weapon-related uses of AI in military systems (e.g., for data analytics). A recent study conducted by the United Nations Institute of Disarmament Research (UNIDIR) distinguishes weapon-specific uses of AI from non-weapon related uses, referring to them respectively as "downstream" and "upstream" tasks.[4] Both weapon-specific and non-weapon-related tasks enable military capabilities, such as command and control, intelligence, surveillance, and reconnaissance (ISR), communications, environmental monitoring, missile defence/early warning, and logistics.

While weapon-specific and non-weapon-related applications of AI may seem disconnected at first glance, military decision-making involves a continuum of activities. Various tasks are, in essence, interconnected in a web of systems of systems. For instance, target selection or attack execution can only ensue after initially searching for, detecting and identifying a threat through ISR capabilities. This means that the decision to use force involves input from both weapon-specific and non-weapon-related tasks.

The interplay between weapon-specific uses and non-weapon-related uses of AI within the spectrum of the military decision-making process has not received sufficient attention from the expert community. There is a need to map out the dependencies between both types of tasks in order to understand potential cascading impacts. In the long run, dependence on AI-enabled systems across the military decision-making cycle may result in new ***unknown*** vulnerabilities.

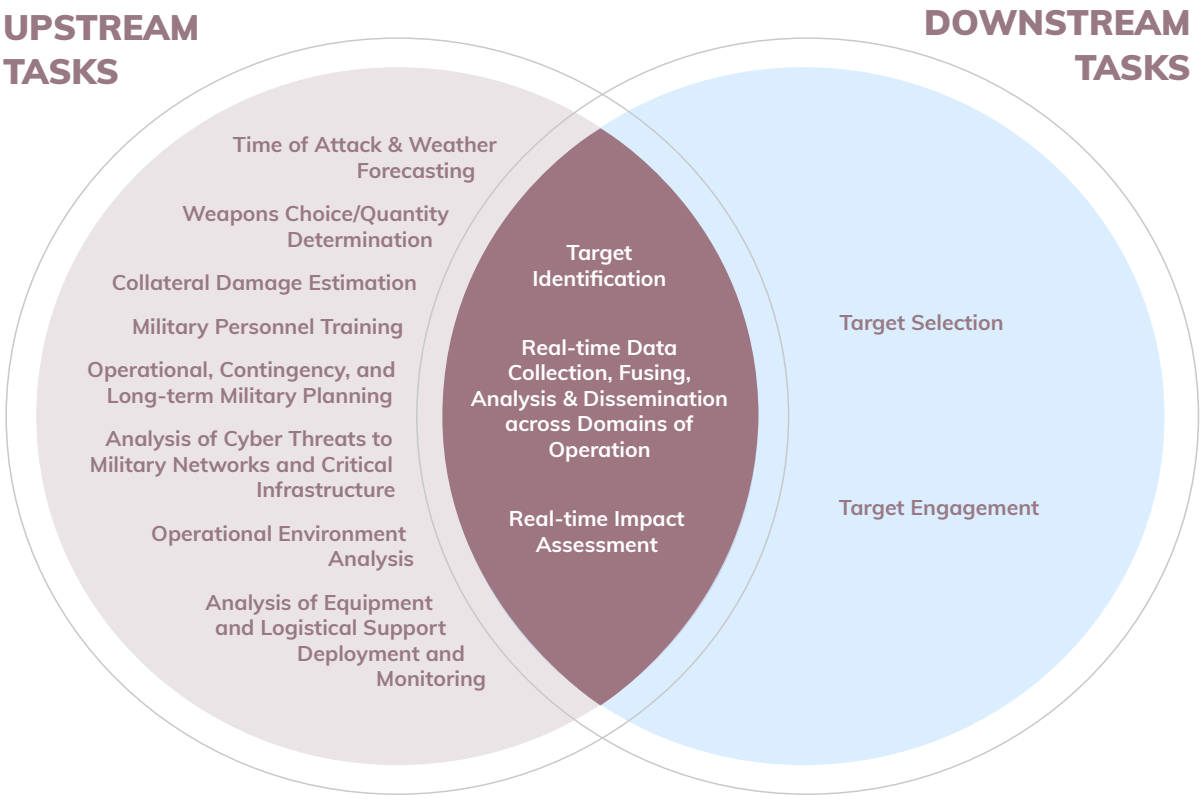
1 Masuhr, Niklas. "AI in Military Enabling Applications." CSS Analyses in Security Policy, Center for Security Studies (CSS), ETH Zürich, October 2019, p.4.

2 For a detailed overview of the tasks, decisions, and actions each level of command involves, see Ekelhof, Merel, and Giacomo Persi Paoli. "The Human Element in Decisions About the Use of Force," March 31, 2020.

3 Stuart Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, 4th ed., Pearson, 2021; Menthe, Lance, Li Ang Zhang, Edward Geist, Joshua Steier, Aaron B. Frank, Erik Van Hegewald, Gary J. Briggs, Keller Scholl, Yusuf Ashpari, and Anthony Jacques. "Understanding the Limits of Artificial Intelligence for Warfighters: Volume 1, Summary", p.2. RAND Corporation, January 3, 2024.

4 Upstream tasks include a range of activities that occur before target selection and engagement. Grand-Clément S., 2023, "Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain," UNIDIR, Geneva.

Figure 4. Continuum Between Upstream and Downstream Tasks



In instances where a weapon system is granted AI-enabled autonomy to pursue its programmed objective without meaningful human control, it is not foreseeable how such a system would achieve that objective. The dynamic operational environment coupled with algorithms based on machine learning "make it extremely difficult to predict the behaviour of these weapons in real-world settings."[5]

The end purpose of use (also referred to as intent), and whether an AI system would be used in a high-stakes environment or integrated into a weapon-specific process (i.e., targeting and attack), could be part of States' assessment of whether certain use cases of AI should be placed off limits.

For instance, AI-enabled ISR could be used for both early-warning and command-and-control purposes. Although AI-enabled ISR data could help analyse satellite imagery, for instance to identify potential threats (e.g., movement of adversarial forces), it could equally be used for the selection of targets and attack execution. While an AI-enabled ISR capability could be used in both cases, some States may find the latter use (i.e., command and control) more problematic due to the potential consequences associated with such use. In this respect, the lawfulness of the use of AI in weapon systems would depend on, inter alia, the context for which they are designed and used, including the operational environment and the type and nature of targets.

5 Autonomous Weapons, The Risk of Autonomous Weapons.

Whether it is a weapon-specific or non-weapon-related use of AI, it is important while designing AI-enabled systems to identify areas where the system would require human oversight and human control.

Integrating AI into a weapons system may add a new dimension of ***unpredictability*** regarding its intended effects. Issues related to reliability, misidentification of targets, and misclassification of civilians as combatants are examples of the potential pitfalls embedded within complex algorithms. The potential use of AI-enabled weapons systems in populated areas also raises humanitarian concerns. The increased physical and cognitive distance between the commander and the battlefield can lead to a lower threshold for using force. The perceived reduced risk to military forces and civilians may also prompt States to resort to force more quickly than they otherwise would have.

The end purpose of use could be part of States' assessment of whether certain use cases of AI should be placed off limits.

The extensive collection and use of data, as well as surveillance technologies such as facial recognition technology, also raise concerns around how these AI-enabled systems have been developed and whether any inherent bias in the datasets might lead to wrongful action, including disproportionate targeting of people of certain demographic groups (e.g., race or gender). These concerns are also directly linked to potential human rights violations.

Spectrum of Autonomy

Recognizing that there is no agreement among States on the definitions and characterizations of autonomy, below is an assessment of ongoing discussions within the expert community on the spectrum of autonomy.

According to experts, systems can feature a spectrum of autonomy — automatic, automated, and autonomous.[6] ***Automatic*** and ***automated*** systems execute predetermined tasks based on pre-programmed instructions. They share a rule-based paradigm, meaning they operate within a set of established rules and scenarios their programmers anticipate. ***Automatic*** systems rest on a simple, threshold-based design and the outcome of the actions from such a system is highly predictable.[7] ***Automated*** systems differ by the complexity of these rules and may rely on "several variables", but they still lack the ability to make independent decisions beyond the initial programming.[8]

6 Scharre, P. Army of None: Autonomous Weapons and the Future of War, p. 31 First edition. New York; London: W. W. Norton & Company, 2018.

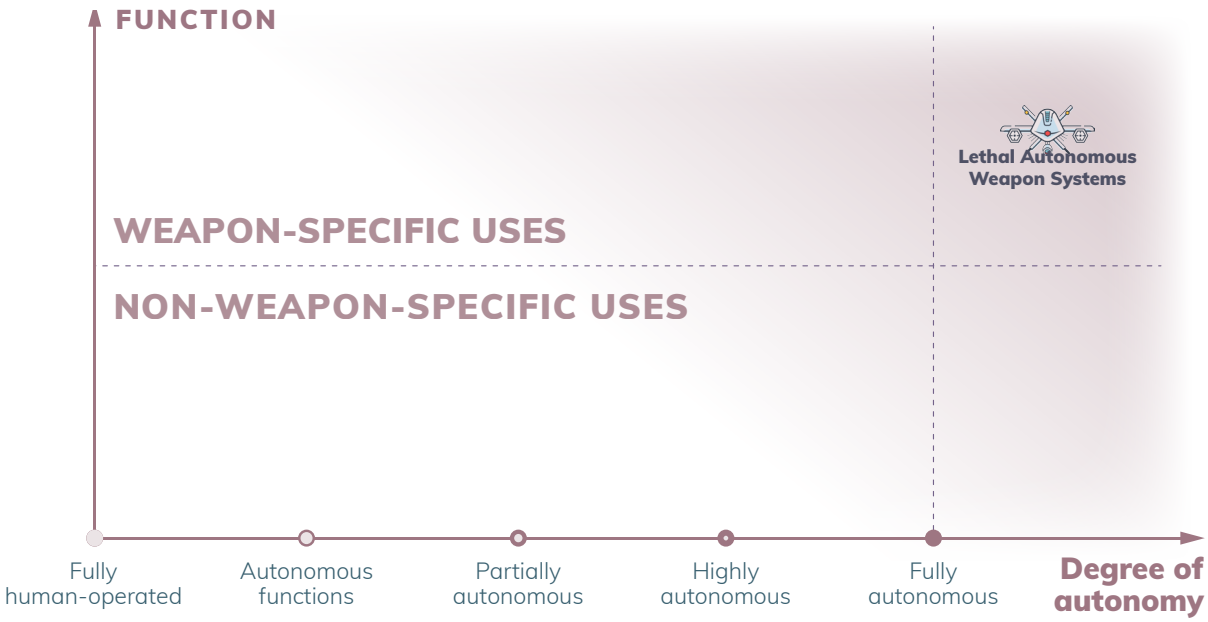
7 Paul Scharre provides the example of old mechanical thermostat as an example of an automatic system. The thermostat reaches a desired temperature or reheats/cools to that desired level. See Scharre P, Ibid. p. 31.

8 Paul Scharre proposes new digital programmable thermostat systems as an example of automated systems, whereby humans pre-program certain conditions for heating/cooling a house, such as whether it is day or night or the time of day/night. See Scharre P., Ibid. p. 31.

Experts indicate that ***autonomous systems*** "exhibit self-direction, self-learning or emergent behaviour" to achieve a goal.[9] The degree of autonomy (e.g., semi-autonomous, fully autonomous) may depend on (a) the level of sophistication of a system, (b) the environment in which it is used (e.g., open versus closed-ended operational environment), and (c) the end goal. For instance, consider an autonomous uncrewed aerial vehicle tasked by a human operator with achieving a specific objective, such as conducting reconnaissance in a designated destination. In such a scenario, the autonomous system may rely on pre-programmed data, such as navigation routes and altitude specifications, to achieve its mission. However, it is not possible for a human operator to predetermine every action for the system due to the changing nature of the operating environment. This is the point where artificial intelligence and machine learning algorithms are of service. By continuously learning from its actions and the operating environment, an autonomous uncrewed aerial vehicle equipped with AI can adapt and make decisions with limited or no human control to accomplish its intended goals. The integration of AI into autonomy, however, makes the actions of autonomous systems inherently unpredictable and their use in certain situations problematic.

9 Ivanova, Ksenia, Guy E Gallasch, and Jon Jordans. "Automated and Autonomous Systems for Combat Service Support: Scoping Study and Technology Prioritisation," p. 2. Defence Science and Technology Group, Land Division, Department of Defence, Australian Government, 2016.

Figure 5. AI and Weapons Systems: Spectrum of Autonomy



There is currently no internationally agreed definition of the levels of autonomy. This diagram only serves illustrative purposes.

AI and autonomy are distinct but related concepts. The former is an overarching technology that encompasses computational methods to undertake tasks typically associated with human cognition, such as vision, speech, writing and reasoning. In recent years, the distinction between the two terms has become increasingly blurry since AI can enhance autonomy, as in the case of lethal autonomous weapons systems.

AI can enable varying degrees of autonomy within specific functions of a weapon system by using (i) pre-defined rules of action based on specific parameters encoded during development or (ii) machine learning techniques that rely on extracting patterns learned from data, going beyond the explicitly programmed input. In the latter context, the autonomous system undertakes actions and makes decisions, at times independently of humans.

AI-enabled autonomous systems are expected to make independent decisions about how to achieve their predefined objectives in dynamic and potentially unpredictable environments. In other words, a human operator may program a strict set of instructions and goals into a system, and the "autonomous system has flexibility in how it achieves that goal."[10]

It is important to note that AI is not a prerequisite for autonomous weapons systems, but when incorporated, AI could further enable such systems. In other words, incorporating AI into autonomous weapons systems is a selective choice and not all autonomous weapons systems incorporate AI to execute tasks. States increasingly rely on AI in weapons systems due to the evolving nature of warfare. For instance, some States may consider AI-enabled weapons systems to be valuable in communication-denied environments, such as where communication systems are jammed by the adversary.

Lethal Autonomous Weapons Systems

While there is no internationally agreed definition of autonomous weapons systems, these could be broadly described as weapons systems that, once activated, select and engage targets autonomously and without further human intervention.[11]

The Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (GGE on LAWS) of the Convention on Certain Conventional Weapons (CCW) has been discussing some considerations on the integration of AI into such weapons systems. Among other things, the GGE on LAWS concluded in 2023 that, "when necessary, States should, inter alia, (a) limit the types of targets that the system can engage, (b) limit the duration, geographical scope and scale of the operation of the weapon system; (c) provide appropriate training and instructions for human operators."[12] During the GGE on LAWS sessions in 2023, many High Contracting Parties to the CCW also raised their positions on technical, ethical and legal considerations relating to the use of LAWS, including matters of explainability, predictability and traceability of such weapons systems, as well as the issue of

10 Scharre, P. Army of None: Autonomous Weapons and the Future of War. First edition. New York; London: W. W. Norton & Company, 2018, p.38

11 United Nations General Assembly, August 1, 2023, Current Developments in Science and Technology and Their Potential Impact on International Security and Disarmament Efforts, A/78/268.

12 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects, 24 May 2023, Report of the 2023 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2023/2.

human control.[13,14] Similar considerations and concerns emerge in relation to the use of AI in the military domain.

The GGE on LAWS also concluded in 2023 that:[15]

a. *IHL continues to apply fully to the potential development and use of LAWS;*
b. *Weapons systems based on emerging technologies in the area of LAWS must not be used if they are incapable of being used in compliance with IHL;*
c. *Control with regard to weapon systems based on emerging technologies in the area of LAWS is needed to uphold compliance with international law, IHL in particular, including the principles and requirements of distinction, proportionality and precautions in attack.*

An existing challenge is how States can ensure that LAWS can be developed and used in compliance with international humanitarian law (IHL).

Human Control

The concept of human control in autonomous weapons systems has been an important subject of discussion within the expert community. This concept speaks to the role of humans throughout the life cycle of such weapons systems.

This concept goes hand in hand with the notion of accountability, as there is a direct relationship between exercising human control and legal responsibility. Still, while States generally agree that it is important to maintain human control over the entire life cycle of the weapons systems, the degree of human control necessary may vary depending on the function.

Some experts, for instance, indicate that machines may better adhere to IHL than humans, indicating that AI-enabled weapons systems could increase precision in targeting. This assertion is problematic from technical and ethical standpoints. While AI may enable greater precision in autonomous weapons systems, the inherent uncertainty on the battlefield may lead to machines going beyond their original goal. The decisions that a machine makes on the battlefield also take place at machine speed, leaving little to no room for human operators to observe, interfere or exercise discretion to terminate actions, once initiated.

13 See for instance, Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects, 3 March 2023, Revised Working Paper Submitted by Austria, CCW/GGE.1/2023/WP.1/Rev.1.

14 See for instance, Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects, 8 March 2023, Proposal for an International Legal Instrument on Lethal Autonomous Weapons Systems, Submitted by Pakistan, CCW/GGE.1/2023/WP.3/Rev.1.

15 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects, 24 May 2023, Report of the 2023 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2023/2.

Experts also argue that "non-human systems do not have the necessary moral judgments to justify their actions in ways that respect victims and thus should not make decisions with such significant ethical implications."[16] This is also why losing control over autonomous weapons systems in armed conflict could lead to catastrophic outcomes. In other words, machines do not have moral agency over the decisions they make.

Taking into account these concerns, the GGE on LAWS concluded in 2023 that States must ensure compliance with their obligations under international law, in particular IHL, throughout the life cycle of weapons systems based on emerging technologies in the area of LAWS. When necessary, States should, inter alia:[17]

a. *Limit the types of targets that the system can engage;*
b. *Limit the duration, geographical scope, and scale of the operation of the weapon system;*
c. *Provide appropriate training and instructions for human operators.*

Any measure taken in this space should also consider the entire life cycle of a weapon system, including pre-design, design, development, deployment, use and decommissioning stages.

As part of meaningful human control, States may also consider regulating human-machine interaction, for instance by incorporating an intervention phase for humans for specific critical tests. Limitations on the targets could include use restrictions in populated areas to minimize civilian casualties and ensure human involvement in the selection of targets and firing of a weapon.

16 Morgan, Forrest E, Benjamin Boudreaux, Andrew J Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. "Military Applications of Artificial Intelligence," p. 34. Santa Monica: RAND Corporation, 2020.

17 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to be Excessively Injurious or to Have Indiscriminate Effects, 24 May 2023, Report of the 2023 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2023/2.

3 International Law and Ethical Considerations

The design, development, deployment and use of artificial intelligence in the military domain opens new questions about how international law applies in this space and whether these systems are being used lawfully.

Once an armed conflict has ensued, *jus in bello*, or international humanitarian law (IHL), governs the conduct of the parties to the conflict. IHL obliges parties to an armed conflict to take into account certain rules and principles that seek to minimize harm and suffering in warfare, including in their decisions to use force. Additionally, international human rights law (IHRL) operates in both peacetime and in war and may therefore apply to elements of an armed conflict.

International Humanitarian Law

IHL provides a delicate balance between "military necessity" and the "requirements of humanity."[1] It "protects people who are not or are no longer participating in hostilities and restricts the means and methods of warfare."[2]

States need to ensure compliance with IHL, including the fundamental principles of necessity, proportionality and distinction, at all times in armed conflict. In certain contexts, AI may even help States to comply with IHL, for instance in the detection of potential breaches of IHL or monitoring of IHL compliance. However, depending on the context, AI-enabled systems may also pose a fundamental challenge for humans to ensure adherence to IHL principles. This challenge mainly arises from not knowing how an AI-enabled system reaches a particular conclusion, how it uses the datasets to reach such a conclusion, or how it learns (if it does) from the given datasets. This means that delegating decision-making functions to AI comes with risks.

International law is addressed to humans. Therefore, humans alone hold the responsibility to comply with international law, including IHL and IHRL. Consequently, determining whether an attack is lawful is a legal consideration that must be conducted by a human. In other words, an AI system cannot do a legal assessment and decide to, for instance, target a building according to such an assessment. Context-specific IHL rules are difficult for a machine to apply in a measurable way, especially when the system is not predictable and the environment is unreliable.

In domains where the target type is military and the operational environment is expected to remain static (e.g., maritime operations) with high degree of predictability regarding the absence of civilians or

1 ICRC, 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.
2 ICRC, "War & Law".

civilian objects, some experts indicate that "international humanitarian law (IHL) problem of distinction (discrimination of combatants from non-combatants) poses fewer problems."[3] Experts also indicate that assessing the proportionality of an attack (e.g., civilian casualties) relative to the anticipated military necessity might be easier in the maritime domain than in other warfare domains.[4] Nonetheless, what "damage" (e.g., damage to the environment, civilians or military objects) means for machines may fundamentally be different from that for humans.[5] Even in such situations, States have the obligation to apply the principles of IHL at all times.

What "damage" means for machines may fundamentally differ from that for humans

In addition to legal considerations, experts have debated the moral significance of human emotions in armed conflict.[6] Human Rights Watch, for instance, indicated that emotions can act as a "safeguard against killing civilians"[7] and that machines lack such tenets.

International Human Rights Law

IHRL applies both in times of peace and during hostilities. The protection of the fundamental rights of individuals is critical, especially considering that AI-enabled systems are likely to be used outside of armed conflict. Moreover, the dual-use nature of AI raises concerns about potential spillover effects and misuse of civilian AI tools in military settings and of military AI tools in non-conflict settings. For instance, military-grade facial recognition software could be employed outside conflict settings, such as in law enforcement or border security operations.

While it is hard to consider all use cases, the fundamental rights of individuals, such as the right to life, the right to security, the right to privacy, and the exercise of fundamental freedoms, should be protected at all times.

AI systems capable of employing lethal force without meaningful human intervention pose a direct threat to right to life, protected under Article 6(1) of the International Covenant on Civil and Political Rights (ICCPR).[8] Increased reliance on algorithmic targeting carries inherent risks of misidentification, miscalculation and escalation, potentially leading to unlawful loss of life. Furthermore, the possibility of AI-enabled systems being utilized for law enforcement purposes, particularly those involving crowd control, threatens the right to liberty and security outlined in the ICCPR. Indiscriminate collection and analysis of personal data may violate the right to privacy and affect the exercise of fundamental freedoms, particularly freedom of expression and peaceful assembly.

Therefore, following a rights-based approach throughout the life cycle of an AI system is important.[9]

3 For a discussion on domains of operations and IHL, see Lucas G., Law, Ethics and Emerging Military Technologies: Confronting Disruptive Innovation,
4 Ibid.
5 Ibid
6 Morgan, Forrest E, Benjamin Boudreaux, Andrew J Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. "Military Applications of Artificial Intelligence," p. 34. Santa Monica: RAND Corporation, 2020.
7 Human Rights Watch and International Human Rights Clinic, "Losing Humanity: The Case Against Killer Robots", 2012, pp. 27, 37.
8 Article 6(1), ICCPR: "Every human being has the inherent right to life. This right shall be protected by law. No one shall be arbitrarily deprived of his life."
9 Raso F. et al., September 25, 2018, Artificial Intelligence and Human Rights: Opportunities and Risks, Berkman Klein Center for Internet & Society at Harvard University, p. 57

Article 36 – Legal Reviews

As provided in article 36 of Additional Protocol I to the 1949 Geneva Conventions, States have a responsibility to determine whether the development of new weapons, means, or methods of warfare they study, develop, acquire, or adopt would be prohibited under international law, in some or all circumstances. Conducting such reviews is a practical measure that all States can undertake to ensure compliance with international law.

While article 36 legal reviews may not apply to all AI applications in the military domain and States are responsible for defining the criteria around such reviews, it is important to ensure that constant care is taken to spare civilians in armed conflict and to always ensure compliance with IHL.

However, there are several challenges with the legal assessment of AI applications in the military domain:

- When AI-enabled systems rely on machine learning tools, the system's performance may change over time through the learning experience. If the system actively learns, its behaviour could become unpredictable unless the learning parameters are clearly outlined and understood by a human operator.[10] Learning capabilities would necessitate ongoing re-evaluations throughout the life cycle of an AI system to ensure continued compliance with IHL principles.
- Different sets of questions would arise from the use of AI for different functions: Is it to execute the targeting of a weapon, or is it to help human operators execute the targeting?[11]

10 See Boulanin V., Verbruggen M., 2017, "Article 36 Reviews: Dealing with the Challenge Posed by Emerging Technologies", SIPRI

11 Ibid, p. 20.

Accountability

AI-enabled systems cannot be held accountable for violations of IHL, as they lack the moral agency and intent that traditionally define legal culpability. In other words, only States and individuals can be held accountable under international law.

The Chairperson's summary of the 2021 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (GGE on LAWS) includes a guiding principle stating that "human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapon system."[12] The guiding principles also captured elements around accountability, stating that "accountability for developing, deploying and using any emerging weapons systems in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control."[13] The report of the session covered these points accordingly.[14]

Traditional accountability frameworks rely on a chain of command where the ultimate responsibility rests with the individual

12 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Chair's summary, CCW/GGE.1/2021/3, Annex 3.

13 Ibid., p. 11

14 Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Report of the 2021 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 22 February 2022, CCW/GGE.1/2021/3.

who ordered the attack. The challenge with AI-enabled weapons systems is to trace the action back to a responsible party — e.g., developers, operators, military commanders or the State deploying them. The responsibility may change in each phase of the life cycle of an AI system.

While it is not yet settled how to attribute and allocate accountability for wrongful conduct by States, the responsible design, development, deployment and use of AI in the military domain is incumbent on all key stakeholders — States, individuals, military actors, operators, designers and the corporate sector, among others.

Further study needs to be conducted around State responsibility across the life cycle of AI-enabled technologies in the military domain.[15]

States can impose contractual terms on defence and private sector companies to conduct legal reviews of their products

The relentless pursuit of technological supremacy in AI by a handful of technology companies also leads to an environment where the imperative to innovate can rapidly overshadow the equally critical need for due diligence. There is a dual responsibility between States and the private sector companies to ensure compliance. Companies involved in the design and development of AI technologies are responsible for ensuring their actions do not contribute to violations of international law. They need to implement necessary due diligence processes in this regard.

These processes should proactively identify, assess, prevent and mitigate potential adverse human rights impacts[16] and be based on human-centric approaches to protect civilians and civilian objects. At the same time, States are responsible for ensuring that corporations under their jurisdiction do not contribute to violations of international law. In this regard, States can, for instance, impose contractual terms on defence companies and/or the private sector to conduct legal reviews of their products, which would complement States' obligation to conduct legal reviews under article 36.

15 For further reading, see, Boutin, Bérénice. "State Responsibility in Relation to Military Applications of Artificial Intelligence". Leiden Journal of International Law 36, no. 1 (March 2023): 133–50.

16 See Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, Office of the High Commissioner for Human Rights, 2012. The Human Rights Council unanimously endorsed these Guiding Principles in its resolution 17/4 of 16 June 2011.

4 Lessons Learned from the Civilian AI Industry

Developments in the civilian AI sector can provide insights into the military domain. Moreover, the design and development of civilian AI applications impact international peace and security in several ways.

- Civilian AI applications can be used in armed conflict by both States and non-State actors. Technology companies may willingly provide access to their products since such access would allow them to test these products in real-world settings. However, the increased involvement of civilian technology companies in armed conflict opens questions about the legal implications such companies face under international humanitarian law.
- Civilian AI models can be repurposed and used for military purposes, such as the surveillance of minorities in armed conflict.
- Civilian AI applications can be misused for malicious purposes — such as to conduct cyberattacks or disinformation campaigns, or to design weapons systems — both in peacetime and in times of conflict. In the future, novel risks may also arise from this risk as civilian AI technology matures.
- How governments and the industry deal with the technical, ethical and governance challenges of civilian AI applications can also shed light on the military domain. For instance, the need for representative datasets to mitigate bias is a common theme across civilian and military domains.

Civilian AI applications share foundational links with their military counterparts: the need for large datasets, computing power, and algorithms. It is, however, important to note that while the baseline testing, evaluation, validation, and verification (TEVV) methods would be similar, AI applications that are designed and used for military purposes would require military grade security standards. Therefore, any use of civilian applications for military purposes should meet the required military-grade standards.

In the civilian sector, open-source datasets and foundational models have been praised for bringing transparency in the field and providing the means for developing countries to develop similar models. Open-source models encourage a collaborative development process by bringing a broad community of developers, ethicists, and users to examine the components of an AI model or dataset. This examination leads to developing more robust and secure AI models. The diversity of perspectives also allows for creating AI systems aligned with ethical considerations, including guidelines and standards. However, once these models are released to the public, regulating how end users employ them becomes challenging.

Conversely, closed-source AI models are often developed by a limited group of contributors; thus, the source code or significant parts of the model may not be released to the public, hindering innovation and oversight. This may lead to perpetuation of bias or errors or to ethical blind spots not being detected at the design stage. Such issues then would surface once the end user engages with the product.

Depending on the business type, some private sector companies currently make their full model available, while others only make their methods or data available. There is no shortcut answer to whether to favour open-source models over closed-source models.

Despite many benefits, there are dual-use risks that open-source models pose to international peace and security. Providing full access to a model may lead to irreversible proliferation whereby the model is retrofitted for malicious uses. Moreover, manipulation techniques (such as jailbreaking[1]) may also cause generative AI models to unintentionally share sensitive information. These concerns led the expert community to propose new ideas. One of those ideas is to build self-destructing models, whereby the model learns from adversarial techniques and prevents being repurposed for malicious purposes while continuing to deliver its regular functions.[2] The concept of "zero trust security" from the cybersecurity field, whereby no user is trusted by default, could also be an example of building open-source models with necessary security architecture embedded in them. This would help prevent illegal and harmful access to the model, not only by end users but also by insider threats.

One of the lessons from the civilian domain is understanding that most foundational AI models (e.g., generative AI) are, at least so far, unreliable when used in a social context. For instance, using AI to determine who might be a criminal or who should get welfare benefits is subject to limitations of datasets used. Moreover, many AI systems fail when tasked with interpreting complex human behaviour due to the dynamic and multifaceted nature of human actions and social contexts. Therefore, expecting AI-enabled systems to consistently deliver accurate responses in social contexts would be deemed challenging.

Civilian Initiatives on Responsible AI

While it is not possible to reference all existing work in civilian initiatives, several organizations actively contribute to the development of common understanding around responsible AI. Some of the knowledge gained in the civilian domain on the responsible life cycle of AI could also be transferable to the design and development of AI systems in the military domain. For instance, most of the civilian initiatives recommend initiating responsible practices at an early stage in the technology life cycle (i.e., from pre-design to design), adopting a holistic approach to AI risks (i.e., beyond direct harmful implications to its end users), and engaging a wide range of stakeholders.[3]

1 Jailbreaking is a process in which the user bypasses the initial software restrictions by prompting the generative AI system, causing the generative AI model to generate potentially harmful content.

2 See, for instance, Henderson P. et al., 2022, "Self Destructing Models: Increasing the Costs of Harmful Dual Uses of Foundational Models," Submitted on 27 Nov 2022 (v1), last revised 9 Aug 2023 (this version, v2),

3 UNODA and SIPRI, "Example standards, tools and practices for Responsible Innovation," Factsheet 5, Responsible AI and peace and security.

Figure 6. Responsible AI Initiatives in the Civilian Spheres: A Selective Overview

Institute of Electrical and Electronics Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems[4]

The Institute of Electrical and Electronics Engineers (IEEE) Global Initiative on Ethics of Autonomous and Intelligent Systems aims "to ensure every stakeholder involved in the design and development of autonomous and intelligent systems is educated, trained and empowered to prioritize ethical considerations."[5] Their efforts include the "Ethically Aligned Design" document,[6] which proposes ethical principles for responsible AI development. The Global Initiative also contributes to developing the IEEE P7000 series of standards, which address specific issues at the intersection of applied ethics and system engineering.[7]

National Institute of Standards and Technology Artificial Intelligence Risk Management Framework (AI RMF 1.0)

The United States National Institute of Standards and Technology (NIST) has developed the AI Risk Management Framework (RMF) in January 2023. This framework is an integral part of responsible development and use of AI systems with the aim of helping organizations to manage risks and promote trustworthy AI systems. The framework categorizes AI harms in three areas: harm to people, harm to organization, and harm to ecosystem. Furthermore, the framework acknowledges challenges in measuring AI risks, such as measuring AI risks in controlled environments versus real-world settings.[8]

Global Partnership on Artificial Intelligence

Launched in June 2020, the Global Partnership on Artificial Intelligence (GPAI) is an international, multi-stakeholder initiative to facilitate multidisciplinary collaboration on the responsible development, use and adoption of AI. In December 2023, 29 GPAI

4 For more information, see "IEEE SA - The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems." Accessed March 12, 2024.

5 See "The IEEE Global Initiative on Autonomous and Intelligent Systems: Key Information, Milestones and FAQs about The Initiative".

6 Institute of Electrical and Electronics Engineers (IEEE), Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 2019.

7 For reference, traditional standards have focused on technology interoperability, functionality, safety, and trade facilitation. For more, see IEEE Standards Association. "AIS Standards." Accessed March 12, 2024.

8 For more information, see NIST, January 2023, "Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1".

members endorsed the GPAI Ministerial Declaration, "reaffirming their commitment to advance safe, secure and trustworthy AI, including, as appropriate, through the development of relevant regulations, policies, standards and other initiatives."[9] The four priority areas under GPAI are (a) developing AI-enabled solutions to "enhance societal resilience to various challenges", (b) enabling AI technologies for low-carbon footprint, (c) using AI for promoting human rights and democratic values, and (d) using AI-enabled tools for global health security.[10]

International Research Center for AI Ethics and Governance

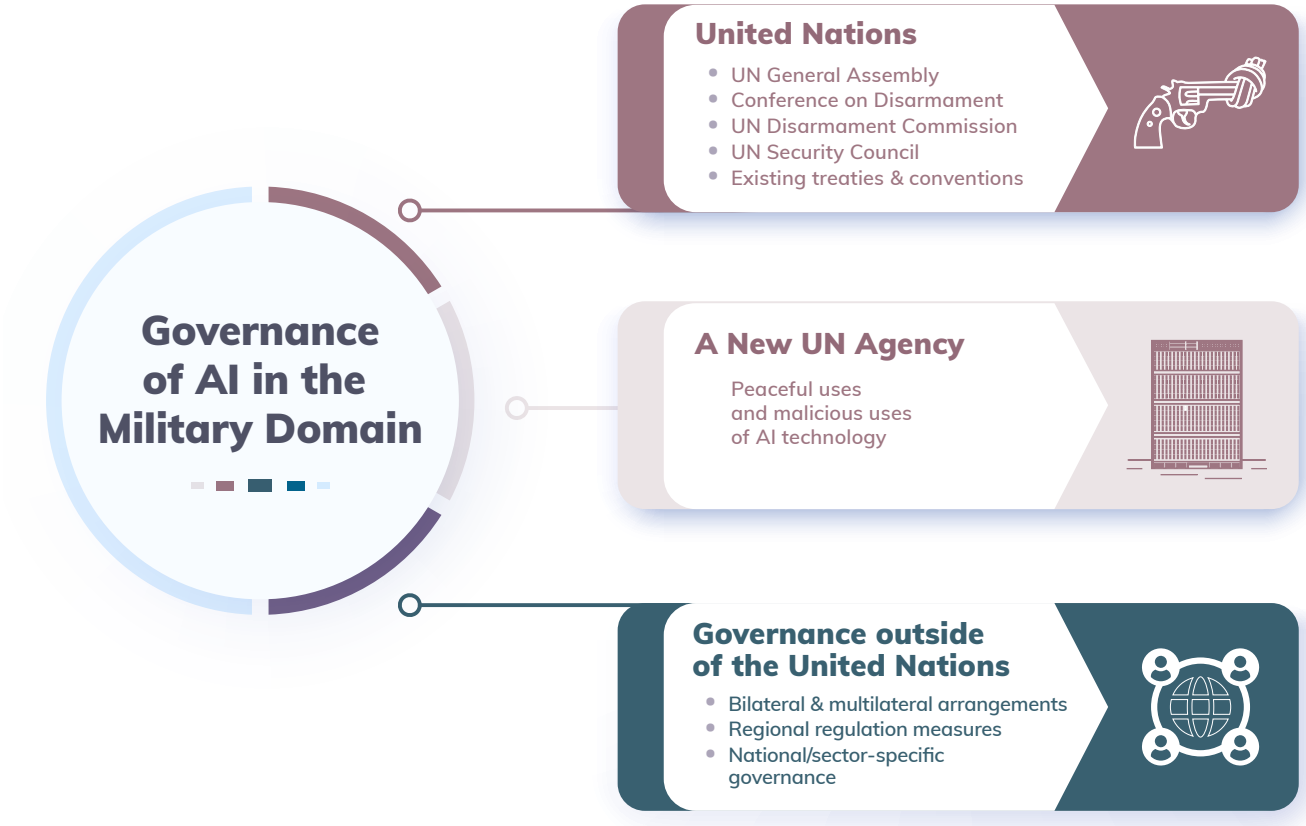
Hosted at the Chinese Academy of Sciences' Institute of Automation, the International Research Center for AI Ethics and Governance aims to ensure that AI promotes good for "humanity and ecology".[11] The Center's work focuses on building global networks to extend "cross-cultural, social and technical collaboration" on AI risks, safety, ethics and governance.[12]

9 See Global Partnership on Artificial Intelligence, 2023, GPAI Ministerial Declaration.
10 The Global Partnership on Artificial Intelligence, 2023, Multistakeholder Expert Group Annual Report.
11 International Research Center for AI Ethics and Governance. "About the International Research Center for AI Ethics and Governance," April 24, 2019.
12 Ibid.

5 Governance Options

States may consider several governance options to address challenges posed by AI in the military domain. Such options may be mutually reinforcing, rather than competing, to prevent the emergence of a single point of failure in AI governance efforts.

Figure 7. Governance Options for Artificial Intelligence in the Military Domain



Option 1: Governance under the auspices of the United Nations in the field of Disarmament

There are various routes available to discuss AI in United Nations disarmament forums. In each case, it is incumbent on Member States to determine the content, scope, mandate and structure of such discussions.

a. United Nations General Assembly

States may consider creating a new agenda item under the seventy-ninth session or future sessions of the United Nations General Assembly (UNGA) First Committee on disarmament and international security.

First Committee resolutions often involve compiling Member States' views, requesting a substantive Secretary-General's report, or forming working groups. Previous working group arrangements have taken the form of open-ended structures where all Member States are invited or groups of governmental experts where the participation is limited.

The UNGA would require a majority ruling to pass such a resolution/agenda item; therefore, carefully crafting resolution language and conducting thorough informal consultations ahead of the First Committee would help penholder State(s) garner support.

b. Conference on Disarmament

AI could be part of an existing agenda item of the Conference on Disarmament (CD). While no existing agenda item currently focuses specifically on emerging technologies, the closest agenda item where AI could be discussed would be "new types of weapons of mass destruction and new systems of such weapons, radiological weapons".

As President of the CD, Germany held an informal meeting at its plenary session on 3 August 2023. Several States delivered statements on this occasion, some of which were published.[1, 2, 3, 4, 5]

Pakistan, in particular, shared its views on the need to set up normative and legal guardrails to mitigate risks of AI "that carry far-reaching security and stability ramifications at the regional as well as international levels."[6] The working paper mainly focuses on AI's impacts on international peace and security, including how it could exacerbate existing risks and create new ones, concepts of deterrence and escalation dynamics, issues around disinformation, and potential manipulation of decision-making.[7] The working paper recognizes the ongoing work within the CCW

1 U.S. Mission to International Organizations in Geneva, Aug 3, 2023, Remarks to the Conference on Disarmament on Artificial Intelligence in the Military Domain (as delivered).
2 Permanent Representation of France to the Conference on Disarmament, Aug 3, 2023, Intervention at the Plenary Session of the Conference on Disarmament – Artificial Intelligence in the Military Domain.
3 Delegation of Japan to the Conference on Disarmament, Aug 3, 2023, Statement by Mr. Umetsu Shigeru, Deputy Permanent Representative of Japan to the Conference on Disarmament, Plenary of the Conference on Disarmament.
4 Conference on Disarmament, Sep 22, 2023, Note Verbale dated 5 September 2023 from Permanent Mission of Ukraine to the United Nations Office and Other International Organizations in Geneva transmitting the statement by the Delegation of Ukraine at the Conference on Disarmament Panel Discussion on Comprehensive Program of Disarmament (Agenda item 6), ways for the responsible development and employment of Artificial Intelligence in the military domain, which was held on 3 August 2023, CD/2376.
5 Delegation of the European Union to the United Nations and other international organizations in Geneva, Aug 7, 2023, Artificial Intelligence in Military Domain Conference on Disarmament – EU Statement.
6 Conference on Disarmament, Note Verbale dated 21 July 2023 from the Permanent Mission of the Islamic Republic of Pakistan transmitting the Working Paper entitled "Addressing the Security and Stability Implications of Military Applications of Artificial Intelligence (AI) and Autonomy in Weapons Systems", CD/2334 (advance copy).
7 Ibid, p. 2-3.

on LAWS and further indicates that "the security implications emanating from the use of AI for military purposes and autonomous weapon systems have direct relevance to several agenda items of the CD."[8]

c. United Nations Disarmament Commission

The United Nations Disarmament Commission (UNDC) is another deliberative body within the United Nations disarmament forums.[9] The UNDC is composed of two working groups and its programme of work runs for three years. While the first working group has continuously focused on issues related to nuclear weapons, the second working group has been flexible and open to any issue area in the field of disarmament.

The UNDC has previously developed principles, guidelines and recommendations. The body was revitalized in its last two sessions, after not being able to agree to a substantive outcome from 1999 to 2017. In the last two sessions, Working Group II respectively focused on "confidence building measures in the field of conventional weapons" and on "transparency and confidence-building measures in outer space activities."[10]

Under the UNDC's newly formed three-year mandate, the second working group will elaborate recommendations on common understandings related to emerging technologies in the context of international security. Under this item, Member States will have the opportunity to share their views on any emerging technology area, including artificial intelligence.

8 Ibid, p. 4.
9 See United Nations Office for Disarmament Affairs, United Nations Disarmament Commission.
10 United Nations General Assembly Official Records, 27 April 2023, Report of the Disarmament Commission for 2023, A/78/42.

d. Existing treaties and conventions

Existing treaties and conventions can incorporate AI into their programmes of work. For instance, in the Programme of Action on small arms and light weapons, States are considering establishing and open-ended technical expert group to discuss recent developments in manufacturing, technology and design. The CCW GGE on LAWS has also discussed AI in its deliberations and could continue to do so, specifically to iron out the links between AI and autonomous weapon systems. Within the Biological Weapons Convention (BWC) framework, several States parties proposed developing a Science and Technology Review Mechanism. In addition, the Treaty on the Prohibition of Nuclear Weapons (TPNW) has established a Scientific Advisory Group.

Option 2: Governance under a Newly Formed Body within the United Nations

The United Nations Secretary-General announced the High-Level Advisory Body on AI (also known as AI Advisory Body) in October 2023.[11] The Body published its interim report on December 2023, with the identification of guiding principles and institutional functions. While the interim report mainly focuses on civilian AI technology, it references AI's "malign and benign" uses for conducting cyberattacks and defending digital networks.[12] It also discusses AI-enabled disinformation and manipulation risks, as well as potential "red lines" of AI, for instance, "in the context of autonomous weapon systems or the broader weaponisation of AI."[13]

11 For more information, see United Nations, Office of the Secretary-General's Envoy on Technology, High-Level Advisory Body on Artificial Intelligence.
12 United Nations, Office of the Secretary-General's Envoy on Technology, December 2023, Governing AI for Humanity.
13 Ibid, p. 10.

The views of States and the multi-stakeholder community differ on whether an overarching governing body on AI, in the form of an Agency or an intergovernmental scientific body, is needed or not. While some States support forming a new body, others indicate the developments of AI cannot be captured under a single governance mechanism. Some States suggest that AI could be governed through a sector-specific approach at the national level. Therefore, whether States would like to consider the malicious uses of AI under such a Body is also an item for further discussion. It is sufficient to say that such deliberations would still not encompass the entire scope of AI's role in the military domain.

Option 3: Governance Outside of the United Nations

There are already ongoing Member State initiatives focusing on artificial intelligence in the military domain. These initiatives can be brought to the United Nations at any point or could stay outside of multilateral forums, should States prefer to do so. These initiatives could also lead to bilateral or plurilateral positions on a specific topic, for instance, around technical elements.

Responsible Artificial Intelligence in the Military Domain (REAIM)

Under the Responsible Artificial Intelligence in the Military Domain (REAIM) initiative, the Netherlands and the Republic of Korea launched a Call to Action during the first REAIM Summit, held in the Hague from 14 to 15 February 2023. This Call to Action invited governments, industry, academia, and international organizations to support the responsible development, deployment, and use of AI in the military domain.[14] The Call has received 57 endorsements as of 1 March 2024.[15] The co-leads are also organizing a second REAIM Summit, which will be held in Seoul on from 9 to 10 September 2024. Regional consultations are also taking place in Chile, Kenya, Singapore and Turkey.

REAIM aims for a balanced approach to cover AI's advantages and the risks associated with the design, development, deployment and use of AI-enabled systems in the military domain. Part of the multi-stakeholder discussion includes international cooperation, coordination, and norm-building. While the international community has yet to determine what can be achieved under the REAIM banner, common understandings on the core concepts of responsible use, minimum guardrails and measures to mitigate risks, confidence-building measures to de-escalate crises, and international cooperation areas to enforce peaceful uses are all part of the discussion.

An essential part of REAIM is the Global Commission on Responsible AI.[16] The Commission will provide recommendations on the responsible life cycle and governance of AI technology.[17] Starting in spring 2024, the Commission is expected to engage in rigorous research, dialogue and analysis to produce actionable guidance.

Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy

During the REAIM Summit in February 2023, the United States also launched

14 For more information, REAIM 2023 Call to Action.

15 For the updated list of endorsing countries, visit "REAIM 2023 Endorsing Countries and Territories".

16 The Hague Centre for Strategic Studies, Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM), Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM) - Commissioners - HCSS

17 Government of the Netherlands, "Call to action on responsible use of AI in the military domain", February 16, 2023.

its Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy. The United States held another launch of this Declaration at the United Nations on 13 November 2023. So far, 54 States endorsed the latest Declaration as of 1 March 2024.[18] This Declaration is composed of 10 measures for the endorsing States to implement.[19] These measures are grouped under three categories: oversight, accountability and assurance. The United States held its first plenary meeting in March 2024; in the next steps, the endorser States aim to conduct exchanges under these three specific categories.

Other AI Initiatives in the Civilian Domain

There are many other international initiatives around AI. Some of these focus on the use of AI for sustainable development, while others focus on AI safety-related elements. Initiatives highlighted below have direct relevance either to international peace and security or to the misuse or malicious use of AI technology, although they represent only a subset of the broader landscape.

Bletchley Declaration

The United Kingdom hosted an AI Safety Summit in Bletchley Park from 1 to 2 November 2023. This Summit captured safety risks and opportunities arising from the use of AI in the civilian domain. While structured around AI safety-related issues, the Summit also captured "potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of [these] AI models."[20] Particular concerns raised in the Declaration that resonate with the international peace and security community include risks in areas such as cybersecurity, biotechnology, and disinformation.[21] Twenty-nine countries were represented at the AI Safety Summit. The next Summits will be hosted by the Republic of Korea, followed by France.[22]

The Global AI Governance Initiative

The Global AI Governance initiative, led by the People's Republic of China, was launched in October 2023. The initiative aims to "enhance information exchange and technological cooperation on the governance of AI." While this initiative focuses on global peace and development as well as international cooperation, it also acknowledges the potential "misuse and malicious use of AI technologies by terrorists, extreme forces, and transnational organized criminal groups."[23] The initiative directly references the discussions at the United Nations, supporting the call to set up an "international institution to govern AI."[24] It is important to mention that not all States support the establishment of a single governance mechanism.

United Nations

The United Nations system has developed various instruments to address AI governance in several policy areas, from guidance to standards. The United Nations Educational, Scientific and Cultural

18 United States Department of State, Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, Bureau of Arms Control, Deterrence and Stability.

19 United States Department of State, "Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, Bureau of Arms Control, Deterrence and Stability", 9 November 2023.

20 Department for Science, Innovation & Technology, Foreign, Commonwealth & Department Office, Prime Minister's Office, 1 November 2023, "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023".

21 Ibid.

22 Reuters, November 1, 2023, "South Korea and France to Host Next Two AI Safety Summits".

23 Ministry of Foreign Affairs of the People's Republic of China, 20 October 2023, "Global AI Governance Initiative".

24 Ibid.

Organization (UNESCO) Recommendation on the Ethics of AI, approved and adopted by 193 Member States in November 2021, is the first global agreement on AI ethics.[25] Drafted by an Ad Hoc Expert Group of 24 experts, its principles include safety and security, human oversight and determination, transparency and explainability, monitoring and evaluation, as well as responsibility and accountability. A companion Readiness Assessment Methodology supports Member States in implementing these principles through a macro-level preparedness assessment that feeds into UNESCO's capacity-building mapping efforts.[26] The International Telecommunication Union (ITU), along with 40 United Nations partners, has also developed standards and the "AI for Good" initiative.[27, 28] These efforts ensure that advancements in AI technology support the Sustainable Development Goals.

Other Governance Frameworks

Several additional civilian frameworks are shaping the responsible development and use of AI systems. Though at different implementation stages, these frameworks collectively aim to mitigate potential risks while promoting positive societal impact.

- The **OECD 2019 AI Principles** provide non-binding principles for responsible AI development.[29] These principles highlight human-centred design, transparency, fairness and accountability, urging member States to implement the principles within their national policies. The G20 AI Guidelines are drawn from these OECD principles.[30]
- The **European Union AI Act**, endorsed by the European Parliament in March 2024, is the first binding worldwide horizontal regulation on AI.[31] Adopting a cross-sectoral approach to regulating general-purpose AI models, the legislation establishes a regulatory framework for using and providing AI systems throughout the European Union and classifies applications according to their level of risk (unacceptable, high, limited or minimal).[32]
- The **Council of Europe's Convention on AI**, still under negotiation, seeks to establish legally binding obligations for AI development and use. It would establish obligations to respect human dignity, the rule of law, and democratic principles in AI development, design, and application.[33]

Inclusive Governance through Multi-stakeholder Engagement

While States are the ultimate authority to decide on the governance of AI within the United Nations, and this section centred on options for States, the inclusion of the multi-stakeholder community — including civil society, academia, industry and AI standardization bodies, among others — into AI governance would help find inclusive solutions. Such inclusion would also enable AI policy and technical communities on AI to come together.

25 UNESCO. "Recommendation on the Ethics of Artificial Intelligence," 2021.
26 UNESCO. "Readiness Assessment Methodology: A Tool of the Recommendation on the Ethics of Artificial Intelligence," 2023.
27 AI for Good, "International Standards for an AI Enabled Future." AI for Good blog, 6 July 2020.
28 International Telecommunication Union. "AI for Good." Accessed 14 March 2024
29 OCDE.AI Policy Observatory. "AI-Principles Overview." Accessed 14 March 2024.
30 "G20 Ministerial Statement on Trade and Digital Economy," Annex, 2019.
31 European Commission. "AI Act | Shaping Europe's Digital Future," March 1, 2024.
32 "The AI Act Explorer | EU Artificial Intelligence Act". Accessed March 14, 2024.
33 Council of Europe. "CAI - Committee on Artificial Intelligence - Artificial Intelligence". Accessed 14 March 2024.

6 Conclusion and Recommendations

This paper addressed the existing and emerging opportunities and risks of designing, developing, deploying and using AI systems in the military domain. Drawing on lessons from the civilian AI sector, it provided insights into key technical, legal and ethical challenges and risks, as well as governance options for the disarmament community. The diplomatic community in the field of disarmament is well-positioned to address the responsible life cycle of AI and to discuss measures to harness benefits while mitigating risks in the context of international peace and security.

AI is an enabler technology, but it could also enable disruptive innovation in the context of international peace and security. It is currently unclear how and at what scale AI is deployed in active conflicts, in which specific military tasks States rely on AI, what impact it has on the battlefield, how it affects (directly and indirectly) the well-being of civilians, and in which areas it facilitates (if it does) the protection of civilians and civilian objects.

The United Nations provides an inclusive platform for any State that wishes to start a multilateral discussion on the responsible use of AI in the military domain. Prior to starting a multilateral dialogue, States should further consider the end goal of initiating such a process: Is it to help develop common understandings and transparency on the use of AI in the military domain? Is it to help reduce tensions, misunderstanding and misperception? Is it to decide on normative frameworks to characterise responsible uses of AI? Is it to identify legally binding measures? While States can choose one consideration over another, all of these areas could be complementary.

There is a dire need to discuss what is considered an *acceptable* and *unacceptable* risk

Moreover, there is a dire need to discuss what is considered an *acceptable* and *unacceptable* risk throughout the life cycle of AI systems. In areas where risks are *unacceptable* due to the extreme or catastrophic consequences to civilians, States may choose to develop normative or legal frameworks.

Developing countries may lack sufficient capacity to elaborate on the military applications of AI. Identifying the urgent capacity needs of developing countries and building targeted programmes to address them at the global level, would help to narrow the AI digital divide.

Recommendations for States

Below is a non-exhaustive list of recommendations and observations for States and other stakeholders to consider for strengthening the work around responsible AI in the military domain.

Recommendations to States on the Governance of AI

- **Initiate** a mechanism to address the governance of AI in the military domain under the auspices of the United Nations. This will ensure inclusive discussions within multilateral forums, engaging not only States that design, develop, deploy and use such technology in the military domain but also States that might be affected by it.
- **Clearly outline** convergences and divergences of lethal autonomous weapons systems and AI by presenting working papers to the CCW GGE on LAWS or engaging within the First Committee. This will help shape both discussions (i.e., on autonomy and AI) to move forward.
- **Establish** national mechanisms to conduct testing, evaluation, validation, and verification (TEVV) of AI applications in the military domain, as well as third-party auditing structures and mechanisms for responsible procurement of this technology from other countries.
- **Design** contracts with AI and defence companies that outline specific roles and responsibilities for those companies to conduct legal reviews of AI-enabled systems as per article 36 of Additional Protocol I of the 1949 Geneva Conventions. While such reviews cannot replace States' obligations, they could provide them with additional vantage points.
- **Establish** mechanisms for developing countries to equitably access peaceful AI applications while preventing proliferation risks and malicious uses by States and non-State actors.

Recommendations to States on Capacity-Building

- **Develop** capacity-building programmes within and outside the United Nations to increase knowledge and expertise on responsible AI in the military domain. These programmes could focus on technical, ethical or legal elements.
- In line with the Secretary-General's recommendation in the New Agenda for Peace policy paper, **establish** necessary national strategies and policies on responsible design, development, deployment and use of AI in the military domain, consistent with the obligations of Member States under international law.[1] The United Nations Secretariat could facilitate a learning programme for developing countries to gain an understanding of the steps and elements involved in the development of national AI defence strategies.
- **Assist** developing countries in conducting legal reviews of new weapons, means or methods of warfare, considering the rapid developments in AI.
- **Provide** appropriate training to operators, military staff and decision-makers to outline both capacities and limitations of AI-enabled systems in the military domain. This could help ensure that humans cognitively engage in the decision-making process and proactively challenge or contest outputs from machines.

1 United Nations, July 2023, New Agenda for Peace, Our Common Agenda Policy Brief 9.

Recommendations for other Stakeholders

Civil society, AI & defence industries, and academia

- **Report** on how civilian AI and defence industries conduct evaluations to assess extreme risks to international peace and security on a voluntary basis. This voluntary reporting could serve as a basis on which States can elaborate to reach military-grade security and safety standards.
- While the industry needs to evaluate and test AI models for extreme risks, governmental actors must **ensure** that these commercial AI applications do not pose risks to national security. The AI industry would not have access to sensitive information to test for certain national security risks. To overcome this problem, AI companies could provide relevant government bodies with early access to their systems to conduct pre-testing before public deployment. This suggestion gained notable attention at the Bletchley Summit in the United Kingdom.
- Civil society and academia can **initiate** track 1.5 to track 2 projects to discuss political, technical, environmental, ethical and legal considerations of AI in the military domain.


9 789210 031745