UNIDIR UNITED NATIONS INSTITUTE
FOR DISARMAMENT RESEARCH

CONFIDENCE-BUILDING MEASURES FOR ARTIFICIAL INTELLIGENCE

A Framing Paper

IOANA PUSCAS

ACKNOWLEDGEMENTS

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. This paper was produced by the Security and Technology Programme, which is funded by the governments of the Czech Republic, Germany, Italy, the Netherlands and Switzerland, and by Microsoft. The author wishes to thank Dr. Giacomo Persi Paoli (UNIDIR) for his substantive input and guidance on this paper.

ABOUT UNIDIR

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

NOTE

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

ABOUT THE AUTHOR

Ioana Puscas is a researcher on AI in the Security and Technology Programme at UNIDIR.

Table of contents

Abbreviations and Acronyms

AI	Artificial intelligence
AWS	Autonomous weapon systems
CBMs	Confidence-building measures
CCW	Convention on Certain Conventional Weapons
GGE	Group of Governmental Experts
IHL	International humanitarian law
ISO	International Organization for Standardization
LAWS	Lethal autonomous weapons systems
NATO	North Atlantic Treaty Organization
OSCE	Organization for Security and Co-operation in Europe
TCBMs	Transparency and confidence-building measures
UNIDIR	United Nations Institute for Disarmament Research
UNODA	United Nations Office for Disarmament Affairs

Executive summary

Artificial intelligence (AI) will shape the future of conflict and warfare in ways that are difficult to predict due to the high uncertainty that characterizes the development and integration of this transformative technology in military capabilities. What is certain is that the increased adoption of AI will introduce new risks to international security that traditional instruments of risk/incident prevention and management may not be adequate to address. Against this backdrop, UNIDIR is launching a project on confidence-building measures (CBMs) for AI, which seeks to explore options that states can consider to mitigate risks and build more confidence and transparency in the development and use of AI in military systems. The project will comprise of two main phases:

1. **Risk mapping**, which will aim to develop a comprehensive overview of the main categories of risks of AI systems, as well as their implications for international security. This evaluation of risks covers a broad taxonomy of risks, such as cybersecurity risks of AI systems, intrinsic risks of the technology (e.g. algorithmic brittleness), or risks related to human–machine interaction.
2. **Exploring possible pathways** for the development of CBMs, which will build on the research findings from the previous phase and will convene multi-stakeholder dialogues with a view to assess realistic options for the development of CBMs.

While more work exists in relation to AI safety and risk management in the context of civilian applications of AI, UNIDIR's project aims at filling a key gap by focusing specifically on military applications and the possible confidence-building framework that could be designed for this unique technology.

Introduction

The military uses of artificial intelligence (AI) and the development and fielding of increasingly autonomous weapon systems (AWS) pose new challenges for military operations and come with risks of misuse, unforeseen incidents and inadvertent escalations. As a blanket ban on AI technologies is both unfeasible and unrealistic, and the continuous innovations in AI methods add further unpredictability on the horizon, it is important for the international community to advance measures for mitigating risks. Policy developments at the national level, and discussions focused on lethal autonomous weapons systems (LAWS) in the Group of Governmental Experts on LAWS (GGE on LAWS) at the United Nations, irrespective of their future outcome (e.g. legally binding instrument or another type of document), cannot address the full range of risks emanating from the military uses of AI.

Confidence-building measures (CBMs) can provide flexible options to articulate rules of the road for future development and deployment of AI systems, and to prevent escalatory consequences at times of crisis or in an armed conflict. While CBMs cannot replace arms control treaties or other binding instruments, they can go a long way in managing risks, clarifying intent, elaborating measures for achieving a goal and creating a baseline of understanding in military actions, including military exercises and operations.

The United Nations Institute for Disarmament Research (UNIDIR) Security and Technology Programme is launching a project that aims to explore possible CBMs for the military use of AI, and methods and steps for developing CBMs for AI. The initial research focus will be on two key areas:

1. **Risk-mapping**: (a) classifying and characterizing risks of the technology and (b) translating those risks into an understanding of their implications for international peace and security. This risk-mapping exercise includes a comprehensive assessment of the risks of AI, as well as scenarios for evaluating the consequences of these risks.

2. **Pathways for CBMs**: understanding what measures states can take to mitigate those risks and exploring pathways for the articulation of CBMs. For example, CBMs may be tailored for the development and testing of the technology in general, for specific use cases, for conduct in an armed conflict or for a combination of these elements. The outcome of this phase of the project will be shaped by multi-stakeholder contributions and dialogues, which will explore lessons learned from other similar processes, and what CBMs options are most suitable, attainable, and realistic for AI.

The project aims to promote an understanding of risks and shared interests to address concerns stemming from the uses of AI across military applications and weapon systems. This is particularly timely and relevant as an emphasis on an “AI arms race” in current narratives promotes the view of a “closed, opaque and competitive culture aimed at ‘winning the race,’ which may ultimately lead to a ‘race to the bottom,’ with the premature – and unsafe – deployment of AI applications (...).”[1] A starting point for the elaboration of CBMs must begin with recognizing shared interests and pivoting the narrative to “a more positive and constructive aim, such as achieving a more reliable, better integrated and more easily explainable technology.”[2] With this project, UNIDIR fills a gap in existing discussions and initiatives related to the governance of AI, by supporting an open exchange between

[1] Giacomo Persi Paoli et al., “Modernizing Arms Control: Exploring Responses to the Use of AI in Military Decision-Making,” UNIDIR, 2020, 91, https://unidir.org/publication/modernizing-arms-control.

[2] Ibid.

stakeholders in order to foster shared understandings and articulate common goals to address the risks of AI.

This paper serves as a framing paper for this new project, introducing its **objectives**, identifying **limitations in existing governance approaches**, and presenting a **possible roadmap for a future elaboration of CBMs**. The scope of the project is deliberately broad at this stage, covering military uses of AI in general and allowing the final outcome to be shaped through views and inputs from multiple stakeholders.

1. Confidence-building measures: overview of the concept

Confidence-building measures (CBMs) in the field of arms control and conflict prevention refer to "planned procedures to prevent hostilities, to avert escalation, to reduce military tensions, and to build mutual trust between countries."[3] CBMs are voluntary and flexible tools to enhance trust and can be unilateral, bilateral or multilateral.[4]

The concept of CBMs rose to prominence during the Cold War and following a series of steps taken by the United States and the Soviet Union to expand their channels of communication, to de-escalate tensions and to make the risk of inadvertent conflict less likely. The concept entered diplomatic language at the 1975 Helsinki Conference on Security and Co-operation in Europe. The "'first generation' of European CBMs" covered exchanges of information, notification and observation of military activities between the 35 participating states.[5]

The voluntary and non-binding nature of CBMs has imposed limitations on implementation and verifiability, but that same characteristic has, at times, provided an incentive for participation and compliance. CBMs do not come risk free, and common challenges include selective implementation, deception or change of course following changes in domestic politics.[6] However, when underscored by strong incentives and shared goals, **CBMs can be an effective tool to adjust inaccurate perceptions, avoid misunderstandings and, over time, stabilize regional and bilateral relations**.[7] CBMs do not address the root causes of conflict and need not, as a rule, limit the national development and adoption of AI technology.

Examples of policy domains that have seen a proliferation of CBMs in recent years are **outer space security** and **cybersecurity**. In the domain of outer space, in 2013, a group of governmental experts established by the General Assembly produced a consensus report which laid out recommendations of voluntary measures to enhance trust and reduce misperceptions and miscalculations (see Box 1).[8]

Box 1. *Characterization and purpose of CBMs as described in the 2013 Report of the Group of Governmental Experts on Transparency and Confidence-Building Measures in Outer Space Activities:*[9]

- CBMs enhance clarity of intentions and create conditions for establishing a **predictable strategic situation** in the economic and security arenas.
- In general, there are **two types of CBMs**: CBMs dealing with **capabilities** and CBMs dealing with **behaviours**.
- CBMs developed in a **multilateral framework** are more likely to be adopted by the international community.

[3] UNODA, "Military Confidence-Building," https://www.un.org/disarmament/cbms/.

[4] Persi Paoli et al., "Modernizing Arms Control," 28.

[5] Marie-France Desjardins, "In Search of a Theory: Developing the Concept," *Adelphi Papers* 36, 307 (1996): 7, https://doi.org/10.1080/05679329608449406.

[6] Marie-France Desjardins, *Rethinking Confidence-Building Measures. Obstacles to Agreement and the Risks of Overselling the Process*, Adelphi Paper 307, 2014 ed. (Abingdon and New York: Routledge, 2014), https://books.google.ch/books?id=6xCgBAAAQBAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

[7] UNODA, "Transparency and confidence building," https://www.un.org/disarmament/convarms/transparency-cbm/.

[8] General Assembly, "Transparency and Confidence-Building Measures in Outer Space Activities," Report of the Secretary General, A/72/65, 16 February 2017, https://undocs.org/Home/Mobile?FinalSymbol=A%2F72%2F65&Language=E&DeviceType=Desktop&LangRequested=False.

[9] General Assembly, Report of the Group of Governmental Experts on Transparency and Confidence-Building Measures in Outer Space Activities, A/68/189, 29 July 2013, https://undocs.org/Home/Mobile?FinalSymbol=A%2F68%2F189&Language=E&DeviceType=Desktop&LangRequested=False.

In the field of **cybersecurity**, CBMs have developed both at the United Nations level (with the work of the UN GGE on Developments in the Field of Information and Telecommunications Technologies in the Context of International Security[10]) and across regional organizations, notably at the Organization for Security and Co-operation in Europe (OSCE), which elaborated two sets of CBMs: the first, in 2013,[11] and the second, in 2016.[12]

[10] General Assembly, Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, A/65/201, 30 July 2010, https://undocs.org/Home/Mobile?FinalSymbol=A%2F65%2F201&Language=E&DeviceType=Desktop&LangRequested=False.

[11] OSCE, Decision No. 1106, "Initial Set of OSCE Confidence-Building Measures to Reduce the Risks of Conflict Stemming from the Use of Information and Communication Technologies," PC.DEC/1106, 3 December 2013, https://www.osce.org/files/f/documents/d/1/109168.pdf.

[12] OSCE, Decision No. 1202, "OSCE Confidence-Building Measures to Reduce the Risks of Conflict Stemming from the Use of Information and Communication Technologies," PC.DEC/1202, 10 March 2016, https://www.osce.org/files/f/documents/d/a/227281.pdf.

2. Roadmap for confidence-building measures for artificial intelligence

2.1. Context: military uses and taxonomies of risks

AI introduces important opportunities for military operations, with uses ranging from logistics, mission readiness assessments,[13] modelling and mission planning, to improved and expedited processing of real-time battlefield data, target identification, navigation and so on.[14]

As a **general-purpose technology,** an evaluation of the risks of AI lends itself to a multidimensional analysis, which includes an analysis of **classes of technical vulnerabilities** (*why* does the system fail or malfunction?) and a transversal **examination** of the range of capabilities that are impacted when an AI-enabled system fails (*what* is the consequence of the failure? and *how* will those failures manifest?).

Firstly, AI systems are vulnerable to several categories of technical risks. Examples commonly identified in technical literature include the following:

- Cybersecurity risks [15] (caused by malicious intent), which apply to all machine learning systems, both in the training phase and for deployed systems [16]
- Intrinsic vulnerabilities that exist in AI learning systems (which may occur even if the system is not "attacked"), such as challenges of **algorithmic brittleness** [17] or **problems of (mis)specification**[18]
- Risks that stem from challenges of human–machine interaction and human–AI teaming [19]

These vulnerabilities are typically studied in a fragmented manner, as different technical communities evaluate risks in their fields of expertise. This project aims to conduct a comprehensive survey of risks, which is critical going forward as it provides the policy community with a broad and holistic understanding of where challenges lie and how various technical risks are connected.

Secondly, because the uses of AI cover a broad range of applications, including in decision-support systems and in autonomous functions (navigation, intelligence, targeting etc.), the consequences of the technology's failure or its suboptimal performance can manifest across capabilities and can have different levels of impact. In addition to a technical survey of risks, the research will explore realistic scenarios to illustrate likely

[13] See Peter Schirmer and Jasmin Léveillé, "*AI Tools for Military Readiness,*" *RAND,* 2020, https://www.rand.org/pubs/research_reports/RRA449-1.html.

[14] For an overview of current uses, see Forrest E. Morgan et al., "Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World," RAND, 2020, https://www.rand.org/pubs/research_reports/RR3139-1.html.

[15] Peter Eckersley, "The Cautious Path to Strategic Advantage: How Militaries Should Plan for AI," Electronic Frontier Foundation (October 2018), 9, https://www.eff.org/files/2018/10/12/the_cautious_path_to_strategic_advantage_how_militaries_should_plan_for_ai_v1.1_0.pdf.

[16] Andrew Lohn, "Hacking AI. A Primer for Policymakers on Machine Learning Cybersecurity," Center for Security and Emerging Technology, December 2020, 5, https://cset.georgetown.edu/publication/hacking-ai/.
The CIA triad of risks in cybersecurity (confidentiality, integrity, and availability) applies to all machine learning systems, and attacks on these systems, despite their complexity, are in fact recognized to be easier and to require less expertise than designing the model itself.

[17] Mary L. Cummings, "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings," *AI Magazine* 42, 1 (Spring 2021): 7–8, https://ojs.aaai.org/index.php/aimagazine/article/view/7394.

[18] Tim G. J. Rudner and Helen Toner, "Key Concepts in AI Safety: Specification in Machine Learning," Center for Security and Emerging Technology, December 2021, 4, https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-specification-in-machine-learning/.

[19] See National Academies of Sciences, Engineering and Medicine, *Human-AI Teaming: State-of-the-Art and Research Needs* (Washington, DC: National Academies Press, 2022), https://doi.org/10.17226/26355.

consequences of the technology's failures or misuses.

For example, AI systems may misassign routes and equipment distribution to troops *(logistics failure)*, misclassify military targets *(targeting failure)*, misguide an uncrewed aerial system *(navigation failure)* or mistakenly prompt an aerial defence system to fire at an object that is not a real incoming threat *(defensive system failure)*. The security consequences of these accidents and malfunctions will vastly differ. By way of illustration, a tactical mistake in mission planning need not lead to increased tensions between parties to a conflict, but a drone accidentally crossing into the airspace of a neighbouring country at a time of heightened tensions may have much graver consequences. Or if the image classification model in a combat drone misclassifies civilian buses as military vehicles and sends the wrong information to the remote command, the consequences can be devastating. The magnitude of the risks would be greatly exacerbated in case of the use of AI in nuclear operations and in nuclear launch platforms.[20]

Unpacking the risks of military AI requires an in-depth study and interdisciplinary approach, and it is the first important step in elaborating CBMs.

2.2. Risk multipliers

Even when AI systems work properly, and are consistent with human intentions, there are other risk factors to consider, which stem from the **cumulative effects** that AI can have in the context of warfare. Military interactions below the threshold of war (such as in disputed territories) rarely, in fact, escalate further as it is usually recognized that human commanders will likely exercise some flexibility and seek off-roads from war.[21] Outputs of AI systems may, however, show less flexibility, either by triggering an automatic kinetic response or by simply making decisions which, although not a result of malfunction, can be very different from those of an experienced commander.[22]

Furthermore, competitive pressures may lead to such risks being exacerbated by **expedited technology adoption**. In order to gain technological advantage, even at the risk of accidents, militaries could adopt an immature technology too quickly.[23] Shortcuts in testing and evaluation can increase the risks of accidents and inadvertent escalation (when intentional actions unintentionally cause escalation by an adversary).[24] In tense situations, the unintended responses from an autonomous system can spiral into further escalation.

2.3. Addressing risks: existing approaches

At the **national** level, an increasing number of states have begun to address the present and anticipated risks of AI through the elaboration of AI strategies and principles for the safe and robust deployment of AI systems.[25] Some documents are specific to defence,[26] signalling a commitment to pre-emptively

[20] Michael C. Horowitz and Paul Scharre, "AI and International Stability: Risks and Confidence-Building Measures," *Center for a New American Security*, January 2021, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

[21] Ibid., 8. Michael C. Horowitz and Lauren Kahn, "Leading in Artificial Intelligence through Confidence Building Measures," *The Washington Quarterly* 44, 4 (2021): 94.

[22] Horowitz and Kahn, "Leading in Artificial Intelligence," 94.

[23] Horowitz and Scharre, "AI and International Stability," 7.

[24] Horowitz and Kahn, "Leading in Artificial Intelligence," 93–94.

[25] Managing risks is, strictly speaking, not a new concern insofar as states have used AI-enabled systems and software in varying degrees during the past decades, and with established certification processes in place to ensure functional safety. Recent advances in AI and machine learning, however, have introduced new categories of risks, including risks of unpredictability, and new challenges for human control.

[26] Examples include, in chronological order: the US DoD's AI Strategy (June 2018), France's report of the AI task force on "Artificial Intelligence in Support of Defense" (September 2019), Australia's technical report of the Defence Science and Technology Group "A Method for Ethical AI in Defense" (2020/ 2021), the UK MoD's Defense Artificial Intelligence Strategy (June 2022). It should be noted that not all defence-specific documents reflect, at the time of writing, official government positions, and some are the result of designated task forces but without an official adoption at the government level. However, their content is likely aligned with future government approaches.

address concerns related to the military uses of AI, as the development and integration of AI into the battlefield accelerates. Despite semantic differences, national AI strategies to date generally stress the importance of safety, robustness and traceability of AI systems, and the importance of building law-compliant systems.

A defence-specific strategy at the **regional** level was adopted by the North Atlantic Treaty Organization (NATO) in October 2021. The strategy lists six principles for the responsible development and deployment of AI in defence within the Alliance: lawfulness, responsibility and accountability, explainability and traceability, reliability, governability, and bias mitigation.[27]

At the **multilateral** level, the articulation of principles for the deployment of AI-based systems has thus far been promoted in the form of general and cross-thematic applicability with the United Nations Educational, Scientific and Cultural Organization (UNESCO) Recommendation on the Ethics of Artificial Intelligence.[28]

In the context of international humanitarian law (IHL), states have discussed the implications of LAWS in the GGE on LAWS, in the framework of the Convention on Certain Conventional Weapons (CCW), since 2013 (formally since 2017). Even though no legally binding document has been agreed thus far in the process, the discussions have advanced member states' understanding of various aspects of autonomy and human control in weapons systems. The adoption of the 11 Guiding Principles in 2019, while not international norms, established an important framework for discussing future developments in autonomous weapons, geared towards deployment that is compliant with IHL.

At the **industry** level, several leading AI companies have already initiated bottom-up governance initiatives, such as through the development of internal codes of ethics, standards and AI principles. As with many national documents, these AI principles typically emphasize values such as trustworthiness, safety and robustness. Explainability has been another recurrent theme.

2.4. Elaborating confidence-building measures

2.4.1. Limitations in existing governance approaches

Addressing the risks of AI to international peace and security needs a broad and multi-stakeholder approach, fit for the wide array of risks introduced by the use of AI.

The elaboration of national AI doctrines can be considered as an early CBM insofar as it signals to other states a clear goal to develop and field robust and safe AI systems. Similarly, the discussions in the GGE on LAWS have created, de facto, a "soft proto norm" on weapon autonomy, as no state can now contemplate autonomy in the critical functions of a weapon without being pointed to the concerns and debates in the scientific community and at the United Nations.[29]

However, there are limitations in existing processes, which warrant the need for additional tools to address risks.

[27] NATO, "Summary of the NATO Artificial Intelligence Strategy," 22 October 2021, https://www.nato.int/cps/en/natohq/official_texts_187617.htm

[28] While a regional piece of legislation, the European Union AI Act from 2021 is also expected to have global effects insofar as it sets standards for categories of AI systems, including high-risks systems. However, the Act explicitly excludes AI systems that are exclusively developed or used for military purposes. See European Commission, "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts," 21 April 2021, https://artificialintelligenceact.eu/the-act/. Other generic **standard-setting documents** have been developed in multiparty frameworks such as the joint ISO/IEC 23053:2022 *Framework for the Development of Artificial Intelligence (AI) Systems Using Machine Learning (ML)*, https://www.iso.org/standard/74438.html. While not specific to military applications and not legally binding per se, such standards are influential across industries.

[29] Frank Sauer, "Autonomy in Weapons Systems: Playing Catch up with Technology," ICRC Blog, 29 September 2021, https://blogs.icrc.org/law-and-policy/2021/09/29/autonomous-weapons-systems-technology/.

At the **national** level, despite a recent proliferation of AI documents by a group of states, the majority of states have not issued national AI positions and strategies. Furthermore, few states have *defence*-specific AI strategies that clearly spell out plans for the development or procurement of AI systems for defence purposes, and fewer still have developed implementation strategies with attendant action plans across AI supply chains.

An added challenge is that the strength of national documents could be diminished if other states do not share the same values and standards of safety and ethics, and especially as the overall development of AI continues in a belligerent and highly competitive environment, which may exert pressures for fast deployment of AI technologies. Moreover, even when developed with the highest standards of robustness and safety, the malfunction or failure of AI systems during use in an international conflict can still have unforeseen consequences.

In the **multilateral** domain, discussions in the GGE on LAWS forum tend to focus narrowly on *critical functions* in AWS and have remained highly divisive on all regulatory structures. Provisions on risk assessments and mitigation[30] are part of current discussions as a matter of principle, and as applied to weapons systems, but the formulation remains general and the implementation entirely at the discretion of states.

2.4.2. A pathway for developing confidence-building measures

As AI technologies are making their way into the battlefield, CBMs can serve as low-risk initiatives or to reduce the likelihood of a worst-case scenario.[31]

Options for agreements or 'rules of the road' norms can build on historical templates (see Box 2) or start from a limited agenda, which can be incrementally layered in time.[32]

States could, for example, focus on fragmented approaches for developing CBMs, such as capability-specific or operations-specific CBMs (e.g. constraining the use of AI in domains of exceptionally high risk, such as nuclear operations[33]), or on broader agendas (e.g. focused on guidelines for the cybersecurity of AI systems). Other stakeholders may also be part of the process, including academic institutions or industry players.

The absence of an international instrument related to the military uses of AI[34] means that the channels to discuss CBMs in a multilateral framework are currently more limited, but such discussions are critical going forward in order to promote a safe deployment of AI technologies.

Box 2. *International Autonomous Incidents Agreement*

Michael Horowitz and Paul Scharre suggested the creation of an **International Autonomous Incidents Agreement**, modelled on the 1972 Incidents at Sea Agreement between the United States and then Soviet Union. This Agreement would focus on military applications of autonomous systems and, like the Incidents at Sea Agreement, it would establish rules of acceptable behaviour, provisions for information sharing about deployments of autonomous systems and channels for consultation at the military-to-military level.[35]

[30] Principle (g) of the 11 Guiding Principles states: "Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems."

[31] See Michael C. Horowitz, Lauren Kahn, and Casey Mahoney, "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?" *Orbis* 64, 4 (Fall 2020): 528–543.

[32] "Stacking" CBMs can, however, only be achieved if the initial steps build trust. Horowitz, Kahn, and Mahoney, "Future of Military Applications," 537–538.

[33] See Horowitz and Scharre, "AI and International Stability."

[34] Unlike in the case of CBMs developed in the context of the Biological Weapons Convention, or in the case of transparency and confidence-building measures (TCBMs) elaborated in the context of outer space security, where there is an Outer Space Treaty.

[35] Horowitz and Scharre, "AI and International Stability," 16–17.

Conclusion

The fast deployment of AI systems into the modern battlefield requires a serious consideration of agreed CBMs to mitigate risks and enhance transparency and predictability in the development and use of AI systems.

Going forward, it is important that discussions about risks mitigation take place at the multilateral level, even as states continue to develop national strategies and standards of ethics for AI.

In this complex and fast-evolving context, UNIDIR’s roadmap for the development of CBMs will take a **risk-centred approach.** This approach will provide a framework for states and other relevant stakeholders to clarify where challenges and concerns are most pressing, and to articulate shared interests.

Bibliography

Cummings, Mary L. 2021. "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings." *AI Magazine* 42 (1): 6–15. https://ojs.aaai.org/index.php/aimagazine/article/view/7394.

Desjardins, Marie France. 1996. "In Search of a Theory: Developing the Concept." *The Adelphi Papers* 36 (307): 7. https://doi.org/10.1080/05679329608449406.

------. 2014. *Rethinking Confidence-Building Measures. Obstacles to Agreement and the Risks of Overselling the Process*. Adelphi Paper 307, 2014 ed. Abingdon and New York: Routledge. https://books.google.ch/books?id=6xCgBAAAQBAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false.

Eckersley, Peter. October 2018. "The Cautious Path to Strategic Advantages: How Militaries Should Plan for AI." Electronic Frontier Foundation. https://www.eff.org/files/2018/10/12/the_cautious_path_to_strategic_advantage_how_militaries_should_plan_for_ai_v1.1_0.pdf.

European Commission. 21 April 2021. "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Act." https://artificialintelligenceact.eu/the-act/.

General Assembly. 30 July 2010. Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, A/65/201. https://undocs.org/Home/Mobile?FinalSymbol=A%2F65%2F201&Language=E&DeviceType=Desktop&LangRequested=False.

------. 29 July 2013. Report of the Group of Governmental Experts on Transparency and Confidence-Building Measures in Outer Space, A/68/189. https://undocs.org/Home/Mobile?FinalSymbol=A%2F68%2F189&Language=E&DeviceType=Desktop&LangRequested=False.

------. 16 February 2017. "Transparency and Confidence-Building Measures in Outer Space Activities." Report of the Secretary General, A/72/65. https://undocs.org/Home/Mobile?FinalSymbol=A%2F72%2F65&Language=E&DeviceType=Desktop&LangRequested=False.

Horowitz, Michael C., and Lauren Kahn. 2021. "Leading in Artificial Intelligence through Confidence Building Measures." *The Washington Quarterly* 44 (4): 91–106.

Horowitz, Michael C., Lauren Kahn, and Casey Mahoney. Fall 2020. "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?" *Orbis* 64 (4): 528–543.

Horowitz, Michael C., and Paul Scharre. January 2021. "AI and International Stability: Risks and Confidence-building Measures." *Center for a New American Security*. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

ISO/IEC 23053:2022. 2022. *Framework for the Development of Artificial Intelligence (AI) Systems Using Machine Learning (ML)*. https://www.iso.org/standard/74438.html.

Lohn, Andrew. December 2020. "Hacking AI. A Primer for Policymakers on Machine Learning Cybersecurity." Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/hacking-ai/.

Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. 2020. “Military Applications of Artificial Intelligence. Ethical Concerns in an Uncertain World.” RAND. https://www.rand.org/pubs/research_reports/RR3139-1.html.

National Academies of Sciences, Engineering and Medicine. 2022. *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, DC: National Academies of Sciences. https://doi.org/10.17226/26355.

NATO. 22 October 2021. “Summary of the NATO Artificial Intelligence Strategy.” https://www.nato.int/cps/en/natohq/official_texts_187617.htm.

OSCE. 3 December 2013. Decision No. 1106, “Initial set of OSCE Confidence-Building Measures to Reduce the Risks of Conflict Stemming from the Use of Information and Communication Technologies,” PC.DEC/1106. https://www.osce.org/files/f/documents/d/1/109168.pdf.

------. 10 March 2016. Decision No. 1202, “OSCE Confidence-Building Measures to Reduce the Risks of Conflict Stemming from the Use of Information and Communication Technologies,” PC.DEC/1202. https://www.osce.org/files/f/documents/d/a/227281.pdf.

Persi Paoli, Giacomo, Kerstin Vignard, David Danks, and Paul Meyer. 2020. “Modernizing Arms Control: Exploring Responses to the Use of AI in Military Decision-Making.” UNIDIR. https://unidir.org/publication/modernizing-arms-control.

Rudner, Tim G. J., and Helen Toner. December 2021. “Key Concepts in AI Safety: Specification in Machine Learning.” Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-specification-in-machine-learning/.

Sauer, Frank. 29 September 2021. “Autonomy in Weapons Systems: Playing Catch-up with Technology.” ICRC Blog. https://blogs.icrc.org/law-and-policy/2021/09/29/autonomous-weapons-systems-technology/.

Schirmer, Peter, and Jasmin Léveillé. 2020. ‘AI Tools for Military Readiness.” RAND. https://www.rand.org/pubs/research_reports/RRA449-1.html.

UNODA. “Military Confidence-Building.” https://www.un.org/disarmament/cbms/.

------. “Transparency and confidence building.” https://www.un.org/disarmament/convarms/transparency-cbm/.

CONFIDENCE-BUILDING MEASURES FOR ARTIFICIAL INTELLIGENCE

A Framing Paper

IOANA PUSCAS

UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

@unidirgeneva @UNIDIR un_disarmresearch