UNIDIR

Confidence-Building Measures for Artificial Intelligence

A Multilateral Perspective

IOANA PUSCAS

Acknowledgments

Support from UNIDIR's core funders provides the foundation for all of the Institute's activities. Work of the Security and Technology Programme on artificial intelligence is funded by the governments of Czechia, France, Germany, Italy, the Netherlands, Norway, Republic of Korea, Switzerland, the United Kingdom, and by Microsoft.

The author wishes to thank Dr. Giacomo Persi Paoli (UNIDIR) for advice and guidance on the project and for the review of this report, Dr. Beyza Unal for her review and suggestions, and UNIDIR experts who shared views at an internal workshop. The final part of this report summarizes the outcomes of a dedicated closed-door workshop; the author also wishes to thank workshop participants for their active engagement and invaluable insights.

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 the publication are the sole responsibility of the individual author. They do not necessary reflect the views or opinions of the United Nations, UNIDIR, its staff members or sponsors.

About the Author

Ioana Puscas (@IoanaPuscas1) is Researcher on artificial intelligence with UNIDIR's Security & Technology Programme.

Acronyms & Abbreviations

AI	Artificial intelligence
BWC	Biological Weapons Convention
CBM	Confidence Building Measure
CCW	Convention on Certain Conventional Weapons
GGE	Group of Governmental Experts
LAWS	Lethal autonomous weapons systems
OEWG	Open-ended Working Group
OSCE	Organization for Security and Co-operation in Europe
UNIDIR	United Nations Institute for Disarmament Research
UNODA	United Nations Office for Disarmament Affairs

Contents

Part I.

1. Background and Purpose of UNIDIR Project

Confidence-building measures (CBMs) have been invoked frequently in recent discussions about artificial intelligence (AI) governance, in different policy forums, and by different actors. Calls for CBMs have emphasized the need for more cooperative frameworks that can help to build more trust or reduce risks of unwanted consequences in the use of AI-enabled systems.

Deliberations in the Group of Governmental Experts (GGE) on lethal autonomous weapons systems (LAWS), which operates in the framework of the Convention on Certain Conventional Weapons, have also covered various aspects related to CBMs for autonomous weapons (which would also include AI-enabled autonomy).[1] States with historically very different positions on the desired

[1] Notably, in the first session of the GGE in 2023, which took place between 6–10 March, CBMs were formally included as a topic in the indicative timetable proposed by the Chair. See Group of Governmental Experts on Lethal Autonomous Weapons Systems, "Food for thought paper – Indicative timetable", https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/Indicative_timetable_-_first_GGE_LAWS_session_2023.pdf. CBMs were also part of the indicative timetable for the first session of the GGE on LAWS in March 2024. See Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, "Indicative timetable of the first 2024 session", https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/Indicative_timetable_-_first_GGE_LAWS_session_2024.pdf.

outcome of the GGE process have expressed support for the development of CBMs. It is nevertheless noteworthy that mentions of CBMs have been consistent with States' dominant positions within the process. A number of States have been careful to highlight that CBMs should not be a substitute for a legally binding instrument,[2] while others consider CBMs an important goal in themselves and not (necessarily) complementary, or adjacent, to a legally binding instrument.[3]

However, the *content*, *format*, and *actionable points* that may be promoted by CBMs in the context of AI have not been discussed at length at the multilateral level. While there is growing consensus and calls for CBMs in this domain, there have been no advanced conversations at the multilateral level on, effectively, how to get there, what CBMs for AI may look like, which actors would be engaged and so on.

With this project, UNIDIR aimed to meaningfully push these conversations forward.

This report presents a framework for conceptual and practical considerations for confidence-building measures for AI, with a particular focus on the advancement of such initiatives in the multilateral domain. Further, it includes preliminary perspectives collected through a workshop and survey with a group of Member States.

This report concludes the second, and final, phase of the UNIDIR project on CBMs for AI. Phase I consisted of a comprehensive AI risk mapping in the context of international security.[4] The risk taxonomy provided a basis for discussing CBMs during the consultation with States, and it can be leveraged in future discussions.

---

[2] This point was, for example, highlighted in the Proposal "Roadmap Towards New Protocol on Autonomous Weapons Systems" submitted by a group of States (available here: **https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fdocuments.unoda.org%2Fwp-content%2Fuploads%2F2022%2F05%2F20220311-G10-proposal-legally-binding-instrument.docx&wdOrigin=BROWSELINK**), and in a Working Paper submitted by the Non-Aligned Movement and other States Parties (available here: **https://documents.unoda.org/wp-content/uploads/2022/08/WP-NAM.pdf**). In its Proposal, Pakistan further highlighted that transparency and confidence-building measures find meaning against the backdrop of legally binging rules, which they complement but do not replace (see CCW/GGE.1/2023/WP.3/Rev. 1, "Proposal for an international legal instrument on Lethal Autonomous Weapons Systems (LAWS)", 8 March 2023, **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.3_REv.1_0.pdf**).

[3] For example, the Working Paper submitted by the Russian Federation simply refers to the role of CBMs in the context of regional and global collective security; see "Working Paper of the Russian Federation 'Application of International Law to Lethal Autonomous Weapons Systems (LAWS)'", 18 July 2022, **https://documents.unoda.org/wp-content/uploads/2022/07/WP-Russian-Federation_EN.pdf**). Türkiye also mentioned the importance of following "a step-by-step approach" and giving priority to CBMs, "in order to create a conductive environment to move forward" [on LAWS], in a statement delivered at the GGE on LAWS on 7 March 2023, available on UN Web TV: **https://webtv.un.org/en/asset/k19/k19n8iayzg** (approx. min. 10:45-15:45).

In a statement in 2023, the United Kingdom raised the point that a legally binding instrument "cannot adequately fulfil the requirement for risk mitigation and confidence building measures". Further, the statement stressed that CBMs do not preclude efforts towards the future emergence of a legally binding instrument and, should such an instrument achieve consensus, CBMs would still be needed to operationalize the instrument. See CCW/GGE on LAWS, United Kingdom Statement, "Item 5 – Topic 6: Risk mitigation and confidence-building measures", 9 March 2023, **https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/UK_Intervention_Item_5_Topic_6_Risk_Mitigation_and_Confidence_Building_Measures.pdf**.

[4] Ioana Puscas, "AI and International Security: Understanding the Risks and Paving the Path for Confidence-Building Measures", UNIDIR, 12 October 2023, **https://unidir.org/publication/ai-and-international-security-understanding-the-risks-and-paving-the-path-for-confidence-building-measures/**.

Scope

The scope of the project applies to AI in the context of international peace and security, in a broad sense. It is not limited to an application area, to military uses only, or specific weapons systems, such as (lethal) autonomous weapons systems.

Structure of the Report

The report first introduces a theoretical and historical overview of CBMs, including a **typology of CBMs**. This theoretical part sets the framework for the discussion, outlining how CBMs have emerged in other domains and the kinds of risks and concerns they address. While the focus is on multilateral processes and multilaterally agreed CBMs, there are also references to CBMs that have emerged out of regional processes.

The next part of the report focuses on CBMs for AI. This section outlines key challenges and concerns that a CBMs process for AI may need to account for, factoring in elements of historical precedent (what can be learned from other domains) and priorities and goal setting. It then puts forward **a list of tentative ideas for CBMs**, covering both narrow and broad measures. This list is not prescriptive and does not necessarily reflect a UNIDIR position on CBMs. Rather, it is an exercise proposed to encourage conversations around concrete CBMs and, in the process, to incentivize early considerations about priorities, to assess comfort levels and conditions for cooperation around CBMs. The final part of this section presents initial reactions from a **workshop and survey with States**, both to the suggested CBMs, and more broadly, to the prospect of developing CBMs for AI.

2. Confidence-Building Measures

2.1. Conceptual Framework

2.1.1. Definition

A standard definition of CBMs in the area of disarmament and arms control describes CBMs as "**planned procedures to prevent hostilities, to avert escalation, to reduce military tension, and to build mutual trust between countries**".[5]

As the concept suggests, CBMs aim to build confidence, which can be achieved and operationalized through a series of actions and initiatives that aim to adjust inaccurate perceptions, to avoid misunderstandings, or to enable further cooperation. In the case of a peace process, for example, CBMs may be aimed at deepening efforts for negotiations without necessarily, or specifically focusing on the root causes of the conflict.[6]

2.1.2. Historical Background

In effect, measures to build confidence, to prevent hostilities or to ease tensions have existed for centuries, yet it was only during the Cold War, in the second half of the twentieth century that **the concept of CBMs** entered diplomatic language following a series of measures agreed at the 1975 Helsinki Final Act of the Conference on Security and Cooperation in Europe. At the time, 35 States agreed to what is known as the 'first generation' of European CBMs, which covered measures related to exchange of information, notification, and observation, on a voluntary basis, of major military activities.[7]

Two further documents consolidated the importance of CBMs during and in the aftermath of the Cold War. First, the 1986 Stockholm Document, which included provisions for verifiable CBMs. Second, the 1990 Vienna Document, which covered measures for immediate risk reduction and for longer-term routine military interactions (e.g., on-site inspections, annual exchanges of military information).[8]

Prior to the 1975 conference and the following initiatives, the United States and the Soviet Union had agreed to measures to permit direct communication between their leaders during a military crisis (the so-called "Hotline Agreement" of 1963, the first bilateral agreement between the United States and the

---

[5] United Nations Office for Disarmament Affairs, "Military Confidence-Building Measures", https://disarmament.unoda.org/convarms/military-cbms/.

[6] Simon J.A. Mason and Matthias Siegfried, "Confidence Building Measures (CBMs) in Peace Processes", in: *Managing Peace Processes: Process related questions. A handbook for AU practitioners*, Volume 1, African Union and the Centre for Humanitarian Dialogue, 2013: 57-77, https://peacemediation.ch/wp-content/uploads/2013/07/AU-Handbook_Confidence-Building-Measures-in-Peace-Processes.pdf.

[7] Marie-France Desjardins, "In search of a theory: Developing the concept", *The Adelphi Papers* Vol 36, Issue 307: Rethinking Confidence-Building Measures (1996), 7, https://www.tandfonline.com/doi/abs/10.1080/05679329608449406.

[8] OSCE Secretariat, "OSCE Guide on Non-military Confidence-Building Measures (CBMs)", 2012, 12, https://www.osce.org/files/f/documents/6/0/91082.pdf.

Soviet Union),[9] a limited but practical measure to limit risks of nuclear confrontation. Later, in 1972, the United States and the Soviet Union signed an agreement to restrict miliary activities in international waters and included specific provisions on notifications of naval exercises (the 1972 "Incidents at Sea" Agreement). These agreements constituted a framework of bilateral CBMs and played a significant part in demonstrating the role of measures to build trust in a complicated strategic environment.

2.1.3. CBMs: Appraisal and Key Characteristics

Confidence-building measures are now a tested instrument in international relations, adopted and applied across a multitude of diplomatic, military, political and arms control processes. From the earlier, narrow focus during the Cold War (mainly centred around 'hard security' concerns, and on reducing risks of surprise attacks), CBMs have been promoted across a wide range of political or arms control processes.

While it is only fair to acknowledge that CBMs have not always been successful or the result not always commensurate to the intended goals, when pursued and promoted by shared interests and a framework of implementation that is feasible, CBMs can be highly effective mechanisms to enhance trust, to foster greater cooperation and adjust inaccurate perceptions.[10] In emerging policy areas, they can set early frameworks for consensus-building over key terminology, and can help to clarify and promote shared understandings of risks.[11]

CBMs can be promoted unilaterally, bilaterally or multilaterally, as well as take various forms depending on the context in which they are applied, such as pre- or post-conflict, or intra- or inter-State.[12] Regional organizations have played a key role in the development of CBMs, and have been able to leverage their understanding of regional contexts as well as access to key stakeholders (e.g., the Organization of American States and the Organization for Security and Cooperation in Europe (OSCE) played important roles in developing CBMs in the digital domain).

There is no universal prescription for identifying CBMs. Measures under the scope of CBMs are situation specific and vary according to the domain area in which they are applied. However, the procedures underpinning CBMs are essential to their effectiveness in practice, and these procedures must be jointly developed and applied to facilitate trust and mutual understanding.[13]

---

[9] The communication links (which comprised both a duplex wire telegraph circuit and a duplex radiotelegraph circuit) were used on several occasions, including in the Arab–Israeli Wars of 1967 and 1973.

[10] United Nations Office for Disarmament Affairs, "Transparency and Confidence-Building", https://disarmament.unoda.org/convarms/transparency-cbm/.

[11] For example, in the cyber domain, one of the measures in the first set of CBMs agreed at the OSCE with Decision No 1106 in 2013 acknowledged "the risk of misunderstandings in the absence of agreed terminology" and decided on voluntary steps to share lists of relevant national terminology, with an added longer-term goal "to produce a consensus glossary". See OSCE Permanent Council, 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] Giacomo Persi Paoli et al., "Modernizing Arms Control: Exploring responses to the use of AI in military decision-making", UNIDIR, 2020, 28, https://unidir.org/files/2020-08/Modernizing%20Arms%20Control%20Final.pdf.

[13] United Nations Office for Disarmament Affairs, "Military Confidence-Building Measures", https://disarmament.unoda.org/convarms/military-cbms/.

Basic Principles of CBMs: Lessons from Outer Space Security

The **Group of Governmental Experts on Transparency and Confidence-Building Measures in Outer Space Activities**, which was established by a General Assembly resolution in 2010, submitted a consensus report to the Sixty-eighth session of the General Assembly in 2013.

While the document discusses the context of outer space activities, it presents a characterization of CBMs that can be informative, more broadly. Notably, the document mentions:

- "there are **two types of transparency and confidence-building measures**: those dealing with **capabilities** and those dealing with **behaviours**"; and
- "transparency and confidence-building measures developed in a **multilateral framework** are more likely to be adopted by the wider international community".[14]

Further, it advances a method for testing the implementation and validation/demonstration of CBMs through a set of key indicators.[15]

	IMPLEMENTATION	DEMONSTRATION
Who	Who should implement the measure?	Who will be able to confirm that the measure has been implemented?
What	What is the measure that should be implemented? Is it clearly identified and understood?	What should be demonstrated to confirm implementation?
Why	What is the value or benefit of performing the measure?	Does a clear understanding of why it is important to be able to confirm or demonstrate implementation exist?
When	When should the measure be implemented?	At what point is demonstration or confirmation performed?
How	**How should the measure be implemented?**	How is implementation of the measure validated, demonstrated or confirmed?

[14] 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, 12, https://undocs.org/Home/Mobile?FinalSymbol=A%2F68%2F189.

[15] Ibid., 15.

2.2. CBMs at the United Nations

CBMs have been a key priority in United Nations disarmament and arms control processes.

The Secretary-General's **Agenda for Disarmament** of 2018 acknowledged the strategic role of CBMs as part of the disarmament toolbox. The definition and characterization of CBMs in the document highlighted the importance of CBMs, as well as their potential value as part of a multi-step process:

> Measures for transparency and confidence-building are often pursued as voluntary means for sharing information with the aim of **creating mutual understanding and trust**, **reducing misperceptions and miscalculations**, **enhancing clarity of intentions**, and ultimately **reducing the risk of armed conflict.** They can serve as a **baseline for the pursuit of legally binding measures**.[16] [emphases added]

Further, the role of CBMs was reiterated in the Secretary-General's "A New Agenda for Peace" of July 2023, which mentioned that trust is "the cornerstone of the collective security system. […] To help reinforce trust, confidence-building mechanisms have been of great value".[17]

The United Nations has led several initiatives over the years in the area of transparency and confidence–building measures, and it has worked on the elaboration of concrete recommendations for Member States.

In the area of conventional arms, it established the Report on Military Expenditures (MilEx) in 1981, and the United Nations Register of Conventional Arms (UNROCA) in 1991.

In its 2017 Report, the Disarmament Commission included recommendations for practical CBMs in the field of conventional weapons, such as **setting up channels of direct communication between Member States** to reduce risks of misunderstandings; **capacity-building and other educational efforts** to promote CBMs; **dialogue** on strategies and policies linked to weapons use, deployment, control, trade and transfer; **advance notification** of major military manoeuvres; and specific **voluntary military constraint** measures.[18]

The General Assembly also mandated the creation of the Repository of military CBMs, an evolving list of tested measures grouped in five large categories:[19]

---

[16] United Nations Office for Disarmament Affairs, "Securing Our Common Future. An Agenda for Disarmament", 2018, 11, https://front.un-arm.org/wp-content/uploads/2018/06/sg-disarmament-agenda-pubs-page.pdf.

[17] United Nations Secretary-General, "Our Common Agenda. Policy Brief 9: A New Agenda for Peace", July 2023, 8, https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf.

[18] General Assembly, "Report of the Disarmament Commission for 2017", A/72/42, https://undocs.org/Home/Mobile?FinalSymbol=A%2F72%2F42.

[19] The table presents a summary of the CBMs published under the Repository. The full list can be viewed here: United Nations Office for Disarmament Affairs, "Military Confidence-Building Measures," https://disarmament.unoda.org/convarms/military-cbms/.

CATEGORY	DISTINCT MEASURES IN EACH CATEGORY / SELECT EXAMPLES
I. Communication and coordination measures	• Information exchange • Communication (e.g., direct communications/hotline) • Troop movement, exercises, and weapon management (e.g., advance notification) • Exchanging and convening personnel
II. Observation and verification measures	E.g., agreement to exchange invitations to observe demonstrations of new weapon systems
III. Military constrains measures	• Troop movement, exercises, weapons (specific restrictions on major military exercises; agreements on acceptable/unacceptable military activities) • Border areas/demilitarized zones (e.g., develop dedicated code of conduct for activities in demilitarized/other zones)
IV. Training and education measures	E.g., teaching CBM approaches in military schools; applying CBM techniques in command and in field exercises
V. Cooperation and integration measures	E.g., establishing joint crisis-management or conflict prevention centres; establishing joint military and/or science and technology research centres/programmes

3. A Typology of CBMs

This section outlines a typology of CBMs, which can be a starting point for future conversations on CBMs for AI. The typology is presented in the following table and is organized in two broad clusters of measures: **measures to take** (transparency CBMs and cooperation CBMs) and **measures to avoid** (constraint CBMs). This categorization draws broadly on frameworks proposed in other regional and multilateral forums where discussions have progressed over the past years, such as in the fields of cyberspace, outer space, or the system of CBMs under the Biological Weapons Convention.

While the distinct formulation of CBMs for each domain is specific and tailored to that domain, this typology aims to provide a **high-level overview** of types of agreed measures across several policy areas, including newer/emerging areas, such as the cyberspace domain. In other words, although there are terminological differences and more granular classifications of CBMs,[20] CBMs generally fit into one of these two clusters of measures.

Furthermore, this typology is not exhaustive, and it does not aim to formulate a canon for CBMs, nor to assess the success of the CBMs in the respective domain. The aim of this typology is to provide a conceptual resource for discussions on CBMs for AI.

As discussions on CBMs for AI start to gain ground, the legacy of existing CBMs can provide a foundational starting point for considering what could be promoted and implemented for AI. The same categories of CBMs described below may also demonstrate where existing frameworks for CBMs may be inadequate for AI or simply not applicable.

Constraint CBMs: Note on Conceptualization

Historically, constraint CBMs, which encourage limits or some form of restraint, have been less frequently promoted by States in the context of multilateral processes (relative to transparency and cooperation CBMs). Constraint CBMs have been especially promoted across bilateral or regional processes and have been often tied to notification requirements. For example, the Stockholm document of 1986 placed constraints on the number of troops involved in miliary activities and tied this measure to requirements of prior notifications.[21]

In theory, some transparency or cooperation CBMs could be said to contain implicit provisions of constraint, meaning that by encouraging one kind of behaviour (e.g., to provide timely notifications of future troop movements in a given area), they implicitly promote restraint on the opposite or adverse behaviour (e.g., pursuing future troop movements in a given area in the absence of prior notifications). In this paper, constraint measures are categorized as distinct CBMs when they are explicitly and originally worded in a language that promotes constraint, either on capabilities or behaviour, and not following post hoc interpretation.

---

[20] For example, the 2013 Report of the Group of Governmental Experts on Transparency and Confidence-Building Measures in Outer Space Activities grouped CBMs under several categories, such as transparency, international cooperation, consultative mechanisms, outreach and coordination. See General Assembly, A/68/189, 29 July 2013.

[21] OSCE, Document of the Stockholm Conference On Confidence- And Security-Building Measures and Disarmament in Europe Convened in Accordance with the Relevant Provisions of the Concluding Document of the Madrid Meeting of the Conference on Security and Co-operation in Europe, 19 September 1986, 13, https://www.osce.org/fsc/41238.

CBMs Typologies

Actions to Take			
	TYPES OF MEASURE		ACTION ITEMS – EXAMPLES
TRANSPARENCY MEASURES	INFORMATION EXCHANGE	ON POLICIES	**Biological Weapons Convention:**[22] CBM E: Declaration of legislation, regulations and other measures.
			Cyber – OSCE 2016:[23] Voluntary **information-sharing on measures** taken by the Participating States to ensure an open, interoperable, secure, and reliable Internet; **meetings of designated national experts** (at least three times/year), within the OSCE framework, to discuss information exchanges and explore appropriate development of CBMs; information-sharing on **national organization, strategies, policies and programmes**; (given the absence of agreed terminology) provision of a list of national terminology related to ICT security, accompanied by an explanation or definition of each term.
			Cyber – UN GGE:[24] **Sharing of information, good practices, lessons/white papers** on existing or emerging threats and incidents; national strategies and standards for vulnerability analysis; **national and regional approaches** to risk management and conflict prevention, including national approaches to classifying ICT incidents in terms of the scale and seriousness of the incident. Information exchange on **national approaches to ICT security**, ICT-enabled critical infrastructure etc., including the legal and oversight regimes under which these operate; sharing of national views on the **classification of critical infrastructure**, sharing of **relevant national policies and legislation**, and **frameworks for risk assessment** and for identifying, classifying and managing ICT incidents that affect critical infrastructure.
			Cyber – OEWG:[25] Sharing national views on technical ICT terms and terminologies.
			Outer Space - GGE:[26] Information exchange on principles and goals of **national space policies and strategies**; publication of information on **national space research and space application programmes**; information exchange on major military outer space expenditure.
			Outer Space – Disarmament Commission:[27] Regular dialogues about national space policies and activities (dialogues could be supported by the United Nations).

[22] United Nations Office for Disarmament Affairs, BWC Confidence Building Measures, https://disarmament.unoda.org/biological-weapons/confidence-building-measures/.

[23] OSCE Permanent Council, 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.

[24] General Assembly, Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, A/76/135, 14 July 2021, https://front.un-arm.org/wp-content/uploads/2021/08/A_76_135-2104030E-1.pdf.

[25] General Assembly, Report of the Open-ended Working Group on Security of and in the Use of Information and Communications Technologies 2021–2025, A/78/265, 1 August 2023, https://digitallibrary.un.org/record/4020967?ln=en&v=pdf.

[26] General Assembly, A/68/189, 29 July 2013.

[27] General Assembly, Report of the Disarmament Commission for 2023, A/78/42, 27 April 2023, https://documents.un.org/symbol-explorer?s=A/78/42&i=A/78/42_7529841.

TRANSPARENCY MEASURES	INFORMATION EXCHANGE	ON ACTIVITIES (AND EVENTS)	**Biological Weapons Convention**[28] • CBM A – Part 1: Exchange of data on research centres and laboratories; – Part 2: Exchange of information on national biological defence research and development programmes. • CBM B: Exchange of information on outbreaks of infectious diseases and similar occurrences caused by toxins. • CBM F: Declaration of past activities in offensive and/or defensive biological research and development programmes. • CBM G: Declaration of vaccine production facilities.
			Cyber – OSCE 2016:[29] Voluntary exchange of information related to ICTs security; responsible reporting of ICT-related vulnerabilities and sharing of associated information on available remedies.
			Cyber – UN GGE:[30] Exchange of national views and practices on ICT security incidents.
			Outer Space – GGE:[31] **Information exchange and notifications** on orbital parameters of outer space objects and potential orbital conjunctions (involving spacecraft to affected government and private sector spacecraft operators); information exchange on natural hazards, and voluntary information-sharing to governmental and non-governmental spacecraft operators of natural phenomena that may cause harmful interference to spacecraft; notification of planned spacecraft launches.
			Outer Space – Disarmament Commission:[32] Share **space situation awareness data and information**, to the extent practicable.
	(RISK REDUCTION) NOTIFICATIONS		**Outer Space – GGE:**[33] **Notifications on scheduled manoeuvres** that may result in risk to the flight safety of other objects; notifications and monitoring of uncontrolled high-risk re-entry events (e.g., re-entry of space objects or residual material); notifications in case of emergency; notification of intentional orbital break-ups.
	CONTACTS AND VISITS		**Outer Space – GGE:**[34] Voluntary familiarization visits; expert visits including visits to space launch sites, invitation of international observers to launch sites, flight command and control centres and other facilities; demonstration of rocket and space technologies.
			Cyber – OEWG:[35] Regular in-person or virtual meetings of Points of Contact to share practical information and experiences on the operationalization and utilization of the global Points of Contact directory.

[28] United Nations Office for Disarmament Affairs, BWC "Confidence Building Measures".

[29] OSCE Permanent Council, PC.DEC/1202, 10 March 2016.

[30] General Assembly, A/76/135, 14 July 2021.

[31] General Assembly, A/68/189, 29 July 2013.

[32] General Assembly, A/78/42, 27 April 2023.

[33] General Assembly, A/68/189, 29 July 2013.

[34] Ibid.

[35] General Assembly, A/78/265, 1 August 2023.

COOPERATION MEASURES	POINTS OF CONTACT	Cyber – OSCE 2016:[36] Nomination of a **contact point** by Participating States to facilitate pertinent communications and dialogue.
		Cyber – OEWG:[37] Consider **nominating a national Point of Contact** at the technical, policy and diplomatic levels, taking into account differentiated capacities. Cyber – OEWG:[38] Tabletop exercises to simulate the practical aspects of participating in a global Points of Contact directory.
		Cyber – UN GGE:[39] Establish **Points of Contact** at policy, diplomatic and technical levels.
		Outer Space – Disarmament Commission:[40] Consider **designating points of contact** to facilitate the notification of potentially affected States of scheduled manoeuvres that may result in risks to the flight safety of space objects of other States.
	DIALOGUE AND CONSULTATIONS	Biological Weapons Convention:[41] CBM D: Active promotion of contacts between scientists, including exchanges for joint research. (NB: *this CBMs was deleted by the Seventh Review Conference in 2011*).
		Cyber – OSCE 2016:[42] **Consultations between Participating States** to reduce risks of misperception and tensions related to use of ICTs; using OSCE as a platform for dialogue, exchange of best practices, awareness-raising and information on capacity-building; activities for officials and experts to support the facilitation of authorized and protected communication channels to prevent and reduce the risks of misperceptions, escalation and conflict, and to clarify technical legal and diplomatic mechanisms; developing mechanisms to exchange best practices.
		Cyber – UN GGE:[43] Continued dialogue through **bilateral, subregional, regional and multilateral consultations and engagement**, with contributions from private sector, academia, civil society and the technical community.
		Outer Space – GGE:[44] **Consultations through bilateral and multilateral diplomatic exchanges, government-to-government mechanisms** (e.g., military-to-military, scientific etc.) to clarify information, to discuss implementation of CBMs, to prevent/minimize potential risks, etc.

[36] OSCE Permanent Council, PC.DEC/1202, 10 March 2016.

[37] General Assembly, Report of the Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security, A/75/816, 18 March 2021, https://undocs.org/Home/Mobile?FinalSymbol=A%2F75%2F816&Language=E&DeviceType=Desktop&LangRequested=False.

[38] General Assembly, A/78/265, 1 August 2023.

[39] General Assembly, A/76/135, 14 July 2021.

[40] General Assembly, A/78/42, 27 April 2023.

[41] United Nations Office for Disarmament Affairs, BWC "Confidence Building Measures".

[42] OSCE Permanent Council, PC.DEC/1202, 10 March 2016.

[43] General Assembly, A/76/135, 14 July 2021.

[44] General Assembly, A/68/189, 29 July 2013.

	CAPACITY-BUILDING	**Cyber – OSCE 2016:**[45] Using OSCE as a platform for dialogue, exchange of best practices, awareness-raising and information on capacity-building; conducting activities for officials and experts to support the facilitation of authorized and protected communication channels [...] to clarify technical legal and diplomatic mechanisms.
		Outer Space – GGE:[46] Bilateral, regional and multilateral capacity-building programmes on space science and technologies (for developing countries).
		Outer Space – Disarmament Commission:[47] Provide assistance and training and transfer technology, data and material (in particular to developing countries).

Actions to Avoid

	ACTION ITEMS – EXAMPLES
CONSTRAINT MEASURES	**OSCE Vienna Document (1999):**[48] Constraining provisions on number of military activities subject to prior notification within three calendar years involving more than 40,000 troops, or 900 battle tanks, etc.; constraining provisions on number of military activities subject to prior notification within one calendar year, and involving more than 13,000 troops or 300 battle tanks, etc.

[45] OSCE Permanent Council, PC.DEC/1202, 10 March 2016.

[46] General Assembly, A/68/189, 29 July 2013.

[47] General Assembly, A/78/42, 27 April 2023.

[48] OSCE, Vienna Document of the Negotiations on Confidence- and Security-Building Measures, FSC.DOC/1/99, Istanbul, 16 November 1999, https://www.osce.org/files/f/documents/b/2/41276.pdf.

Lessons from other domains: Cyber, Nuclear, Outer Space, Biological Weapons/ BWC

The processes underpinning the elaboration of CBMs in other domains, as well as the resulting measures, provide valuable insights that may be carried forward as States begin to articulate CBMs for AI. The following overview derives from comparative views collected at an internal workshop, which convened UNIDIR experts from several disarmament areas, including the cyber domain, outer space, and weapons of mass destruction (biological weapons, nuclear weapons).

Generally, CBMs have enjoyed wide support at the multilateral level, even as States recognize the many challenges in developing and implementing measures that are not enforceable through legal agreements.

The moral and political obligation attached to respecting agreements, though voluntary, has meant that non-cooperative behaviours or non-compliance received stern political reactions, especially when long-standing and decades-long practices were abandoned unilaterally or abruptly (e.g., self-reporting of assets or activities in the **nuclear** domain, such as through the International Atomic Energy Agency). There emerges, in other words, an expectation of compliance, and a political obligation attached to the process even if it is 'voluntary'.

In domains as intangible as **cyber** (arguably opaquer than domains such as bio or nuclear), where it is difficult to understand or monitor national capabilities and activities, self-reporting is essential and the cornerstone of multilateral processes. In that regard, participation in the Open-ended Working Group was in itself considered a CBM (a point that echoes the views of many delegations at the GGE on LAWS) although it is acknowledged that not all States are able to participate in the same way due to limited capabilities.

Self-reporting is also an established practice in the **Biological Weapons Convention** (BWC) domain, through the CBMs forms, but submission of forms has been rather scant and inconsistent. Here too, lack of capacity, including trained personnel, has been an issue; however, States' lack of willingness to participate is recognized as the bigger and more persistent challenge. Further, even as submitted material could be done promptly there are other inherent risks when the information that is submitted may be inconsistent with publicly available information about bio-related activities in a country, when States may decide to declare no major changes despite significant advances in the biotech industry, or simply due to more practical reasons: States may submit incorrect information not because they intend to violate the Treaty but due to factors such as time constraints or miscoordination between national agencies. One or a combination of such factors may alter the overall trust in the process and be counter to its aim, which is to reduce ambiguities and doubts.

A delicate balancing act between transparency, on the one hand, and protecting what States view as the purview of national security interests, on the other, has also characterized the CBMs process in the context of **outer space**. This has resulted in lengthy attempts to

agree on implementation plans and an overall stalled process, which incentivized commercial actors (motivated to protect their own investments) to pursue parallel efforts for defining best practices—in effect, 'unofficial CBMs'.

In the **cyber** domain, the technical community that effectively works on cyber incidents has a history of operationalizing technical standards. To many in this community, the conversations on CBMs at the multilateral level are a rebranding of standards or regulations that have already existed for some time, and in that sense a potential duplication of efforts. Yet, there is broad consensus that bringing States together and promoting CBMs in the multilateral framework carries political benefits and has value in reducing risks of misunderstanding and unpredictability.

The formulation of technical and scientific language that is part of CBMs is important, especially in the context of ever-evolving technical and scientific advancements. This critique was raised on numerous occasions in the **bio/BWC** context where it was highlighted that the CBMs forms have run into the risk of being surpassed by changes in science.

Part II.

1. Confidence-Building Measures for AI: From Concept to Action

To date, some national initiatives constitute, in effect, incipient forms of CBMs. For example, a proliferation of national AI defence strategies in recent years can be seen as a measure to outline high-level principles for the development and use of AI in military contexts. The language across these documents tends to highlight key principles related to safety, testing and responsible use. Though not necessarily framed as CBMs, these documents can provide reassurance that States commit to being responsible actors in how they develop, procure and use the technology.

At the very least, this may signal to other actors a commitment to deploy the technology in a way that would minimize risks of accidents or unintended behaviour and escalation. At the multilateral level, the 11 Guiding Principles,[49] adopted by the GGE on LAWS in 2019, established important baseline principles for the work of the Group. Though non-legally binding, these principles codified key shared understandings about LAWS among the large group of States which are part of the GGE. The Principles remain restrictive in their scope, however, as they refer to a specific class of weapons systems.

In the context of AI, the need for CBMs is prompted by several factors. There remain many critical areas of risks in the use of AI, which cannot be mitigated by national strategies alone. For example, such documents do not per se build more trust among adversaries, they do not build channels of communication or a shared language of risks and concerns, and do not clarify how specific incidents involving AI-enabled systems would be managed. The growing and wide-scale use and adoption of AI, already present across weapons systems and domains of warfare, warrants the need for meaningful conversations at the multilateral level about how to manage the development, adoption and integration of this powerful technology.

Existing CBMs may provide useful lessons on how to articulate the language of CBMs, understand where stakeholders with very different interests have historically found it easier to reach consensus language and articulate shared goals. The intrinsic complexity of AI technologies, however, including AI's scalability and potential for use across diverse domains, may also mean that existing CBMs may not be suited as templates, or only with limited use.

As a **next step**, a conversation about CBMs could take into account a combination of narrow and broad concerns, including:

---

[49] GGE on LAWS, Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2019/3, 25 September 2019, https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf.

Objectives	• What should be the aim of CBMs for AI? Build *confidence* about what exactly? • How can common interests be identified? • What are key priority areas and the risks to address in a CBMs framework?
Process	• How can States begin a process of articulation of CBMs? • How would such a process function and what resources (financial or otherwise) would need to be committed to the process? • Which actors need to be part of the process?
Content	*Phase I of this project introduced a taxonomy of risks of AI, which can be leveraged to advance initial substantive discussions about the content and articulation of CBMs for AI. The taxonomy identified two large clusters of risks: risks inherent to AI technologies (**safety**, **cybersecurity**, and **human–machine interaction** risks) and risks of AI to global security (**miscalculation**, **escalation**, **proliferation** risks).* Discussions about confidence-building measures could begin with an initial consideration of risks, and address questions such as: • What CBMs can address risks of AI technologies? • How can different types of measures (e.g., transparency measures or cooperation measures) address concerns about the risks of AI technology? • What type of CBMs could address miscalculation/escalation/proliferation risks?

2. Food for Thought List: Possible CBMs for AI

Initial discussions could explore concrete ideas for CBMs as well as tracks that States can take moving forward (e.g., early plans of work on how to elaborate CBMs). A conversation over suggested measures is a useful exercise to clarify views and priorities.

Below is a list of possible ideas that States can consider, which are proposed with the aim to facilitate a meaningful dialogue, as well as to provide an opportunity for States to test their level of comfort as to the nature of engagement and collaboration they are ready to initiate, at least in an early phase. The scope of the suggested CMBs includes **measures relevant to international security** and is therefore not limited to military applications only.[50]

The process of articulating CBMs for AI can grow incrementally. States could develop, for example, dedicated or single-purpose measures focused on AI safety (e.g., incident reporting mechanism or database, etc.), which may evolve in time to other forms of exchange and cooperation. Another possibility is that States may view a formalized process to be the most fitting option going forward (see the last suggestion in the list), and thus allow themselves to consider and test the feasibility of specific CBMs within an institutionalized framework.

The list below is purposefully heterogenous: some proposed CBMs are more encompassing than others, while some are rather narrow. Further, among the suggested options, some CBMs, though agreed at the multilateral level, could require operationalization to be restricted to national action plans while others would involve active international cooperation. CBMs regimes afford this level of flexibility.[51]

LIST OF SUGGESTED CBMs	OPERATIONALIZATION AND IMPLEMENTATION OPTION(S)
Promote the elaboration of national AI strategies (for States which have not yet done so)	Publicly commit to elaborate national strategies for AI Voluntarily share best practices on AI strategies elaboration
CBMs for AI safety: support the convening of technical experts to elaborate common definitions and standards for AI safety	Promote exchanges at working level between national experts to discuss key terminology and standards related to safety and security of AI systems (this may include, for example, horizon-scanning exercises to discuss risks in emerging AI technologies and applications); these efforts may draw on existing standards, frameworks, and progress achieved in the civilian domain

[50] Relatedly, mentions of AI across the table refer to applications of AI that are relevant in the context of international security and must be contextualized as such.

[51] In some cases, the chosen method(s) to operationalize a CBM may also change or broaden the nature of that CBM, e.g., a transparency or cooperation measure. States could decide to exercise transparency over a specific process, such as by publicizing relevant information on legal reviews for AI-enabled autonomous weapons, for example. Additionally, they may also initiate international exchanges and capacity-building to bolster the capacity of more States to carry out legal reviews, thus making a CBM focused on legal reviews not only a measure of transparency, but also of international cooperation.

	Initiate new international mechanism(s) for exchanges on safety-related issues (e.g., an intergovernmental forum at the international level, or under the United Nations)
Promote legal reviews for AI-enabled weapons systems	Publicize high-level information on methodologies and steps for conducting legal reviews; this could include information on the process, departments/expert groups involved etc. while protecting all sensitive information Provide funding or sponsor training and capacity-building for experts tasked to review legality of weapons systems
Military-to-military dialogues to exchange on doctrines, rules of engagement	Exchanges could take place at agreed frequency (e.g., annually) to discuss scenarios (e.g., involving AI-enabled autonomous systems etc.) that may result in escalation, to clarify approaches, and to disseminate the outcome of discussions within national structures Exchanges could also discuss and clarify issues emerging from non-weapons use cases, specific risks and mitigation strategies
Promote and fund civilian research on AI safety and cybersecurity	*Unilateral measure*: States could commit to bolstering AI safety research in academic and research centres, and enable civilian leadership of AI safety research (to promote standards of safety and security to be established in open research settings) *Multilateral measure*: promote Track II exchanges between scientific and academic experts (e.g., to share best practices for incident-reporting, red-teaming, etc.)
Application or domain-specific CBMs	Identify top priorities and areas of high/unacceptable risks and agree on key principles and approaches; these could be in the form of politically binding documents, such as joint declarations, and could cover key concerns such as: refraining from using AI in specific contexts or applications, high-risk domains; defining strict boundaries for use of AI technologies, including AI-enabled autonomous weapons
(L)AWS-specific CBM	Code of conduct for autonomous weapons, which could also define 'rules of the road', clarify red lines for autonomous weapons deployment, and ways to avoid unintended escalatory events
Share best practices for AI testing and evaluation	Publicize, on a voluntary basis, best practices and lessons learned
Promote efforts for non-proliferation and prevention of deliberate misuse	*Unilateral*: Strengthen domestic mechanisms for counter-proliferation, which could incorporate risk assessments and mitigation strategies to respond to AI-generated risks *Multilateral*: Convene experts and promote awareness about AI risks across other processes and disarmament bodies (e.g., risks of AI in the field of biological weapons/BWC)
Create a GGE on CBMs for AI	Establish a GGE (e.g., at the General Assembly) to promote routinized exchanges on specific themes and deliberate on the elaboration of CBMs

3. Confidence-Building Measures for AI: Views from States

3.1. Multilateral Perspective: Workshop Summary

The topic of CBMs for AI as well as the measures suggested in Part II, Section 2 were discussed in a workshop with select Member States in end May 2024;[52] the main conclusions are summarized below. UNIDIR convened a multilateral meeting gathering delegates from a wide and diverse range of countries, with representatives of all regional groups.[53]

The views and recommendations expressed during the discussions provide compelling arguments about the relevance and scope of CBMs for AI going forward as well as many considerations that must be heeded as part of the process of developing such measures.

The content of the discussions is grouped under the following categories:

- Relevance and objectives of CBMs
- Regional versus multilateral CBMs
- Dual-use and opportunities of AI
- Role of capacity-building
- Process

Relevance and Objectives of CBMs

Generally, the discussions highlighted that the flexibility of CBMs makes them useful and relevant for AI and can provide a favorable opportunity to start a dialogue (even) before the technology is regulated. CBMs ultimately aim to help States fulfill obligations under the Charter of the United Nations and the multilateral dialogue on the uses and risks of the technology is particularly important to avoid a fragmented or piecemeal approach. The value of regional discussions on CBMs was also highlighted, with examples from other domains, notably cyber (these different levels are elaborated below), though regional and multilateral processes were, generally, rather seen as complementary.

It was also mentioned that CBMs can establish necessary guardrails to respond to the general lack of predictability around the technology, at least in the present context. Appreciating the risks of AI to international security requires a multi-sectoral approach, as AI's likely impact will span conventional weapons and conflict, and WMD and strategic risks. CBMs can be leveraged to respond to a common interest of all members of the international community to achieve 'some certainty' about the technology and its use. Relatedly, another point was raised about the potential role of CBMs in encouraging responsible development of the technology and opening up channels of cooperation at different levels, and that options for collaboration could include elements such as incident

---

[52] The workshop observed the Chatham House Rule. The summary of the discussion in this report makes no attribution to national delegates or Member States.

[53] United Nations, Regional groups of Member States, https://www.un.org/dgacm/en/content/regional-groups.

reporting or some form of mechanism for signalling unexpected system behaviour.

The lack of agreed terminology was also brought up: AI lacks a clear definition, and it does not refer to one discrete technology or capability, which may hinder efforts to develop further measures. In response, examples from other domains were brought up, such as outer space, where the lack of a definition for a space weapon, for example, did not preclude the development of CBMs. Further, the process of developing CBMs in the case of outer space did not follow (or result from) a full use and militarization of outer space, and norms could emerge ahead of definitions shared by all States.

Regional versus Multilateral CBMs

Drawing on examples from the cyber domain, it was noted that CBMs have worked well at the regional level. The question of the appropriate level at which to start the discussion on CBMs (regional versus multilateral) was raised multiple times.

The value of regional (and in some cases, bilateral) CBMs was recognized for significant progress and achievements in other policy domains. Furthermore, in the case of AI, recent initiatives led by select States were mentioned as examples of processes where the United Nations does not have the leading role, which raises the question if the United Nations should be the main forum for these conversations.

It was also reiterated that, historically, CBMs have had a specific connotation in international relations, and that they were commonly agreed upon by adversaries, either through bilateral channels or at the United Nations. Multilateral forums can thus be an important venue to garner broad support for CBMs, yet it was mentioned that the start of a conversation can be especially facilitated once there is a clear understanding of specific use cases.

Nevertheless, the strength of multilateral engagement was broadly appraised for reasons both pragmatic and foundational to the development of CBMs. First, because multilateral dialogue and establishing common norms for security helps address concrete challenges, such as to prevent outliers. In the absence of global norms, malicious actors could exploit more opportune alternatives to, it was observed, “move to other countries that do not have constraints”.

Second, it was highlighted that discussions in smaller groups would be more fragmented and promote local understandings of standards and risk management, rending it “impossible” to later broaden and socialize these as common understandings.

Dual-use and Opportunities of AI

Both the dual-use nature of AI technologies, and the opportunities afforded by AI were mentioned on multiple occasions. In the military domain, for example, the use of AI was credited for enhancing battlefield awareness and potentially the protection of civilians in the context of armed conflict. Outside of military applications, AI plays a critical role for digital transformation and for meeting targets under the Sustainable Development Goals.

With these remarks, the aim was to highlight that conversations about CBMs may lose sight of the technology’s opportunities and overemphasize risks. While discussions about security were deemed important going forward, they must address specific risks, such as malicious uses of the technology, risks that AI brings to certain weapons systems, or how it can transform the nature of conflict.

Role of Capacity-Building

The role of capacity-building received distinct attention, both as part of the process of considering CBMs and for their implementation. It was recognized that the existing disparity in the development and adoption of AI technology, and between States at the leading edge versus those that are not, must be addressed to develop common ground, build trust, and to allow discussion of the technology's *global* implications.

The call to "close the gaps" resonated with several participants, who supported the idea that an inclusive and multilateral process must engage wide participation, including from the Global South and from States that are not developing the technology. This latter argument emphasized that for AI, capacity-building efforts must attempt to go further than what was done in other processes. Moreover, it was added that building expertise and preparedness (through capacity-building) could also help implementation of CBMs, and this is particularly significant as in the long-term, some CBMs may become legally binding rules.

Process

Several points were raised about the process underpinning CBMs, and their future elaboration. A recurrent point was about the importance of dialogue, which could start with a conversation about risks and ways to address them. A related point was about the critical role of transparency: to develop a shared understanding of risks, States must be transparent in how they view the risks of AI technologies, so as to promote a definition of risks that "is the same for everyone".

Other views underlined the importance of scientific input into the process of elaborating CBMs, or potentially taking lessons from other relevant discussions happening in domains such as export controls. Another perspective was that it is difficult to consider CBMs (for AI) through traditional approaches and that the essentially dual-use nature of these technologies brings with it the need for multiple forms of interaction, including with industry.

Finally, it was noted that a framework of CBMs for AI that would distinguish between 'measures to take' and 'measures to avoid'[54] could be useful and that this framing has proven to work in other contexts.

[54] The distinction referred to the framing discussed in the first part of this report, which was initially shared with States as a working paper prior to the workshop.

3.2. CBMs Survey: Acceptability and Impact

The list of suggested CBMs, including the implementation and operationalization options (see Part II, Section 2), were surveyed twice during the workshop: once for "**acceptability**" and once for "**impact**", using scoring values from 1 to 4 (1 – no acceptability, 2 – low acceptability, 3 – moderate acceptability, 4 – high acceptability; and 1 – no impact, 2 – low impact, 3 – moderate impact, and 4 – high impact, respectively).[55] "Acceptability" essentially referred to the political acceptance and feasibility of the proposed measure, and "impact" to the expected positive effect of the measure.

The results of these surveys should be interpreted as illustrative of the consultation with the States that participated in the UNIDIR initiative. While they reflect views of a diverse group of States and regions, they do not represent the perspectives and preferences of the entire multilateral community; a larger participation base may have produced different results.

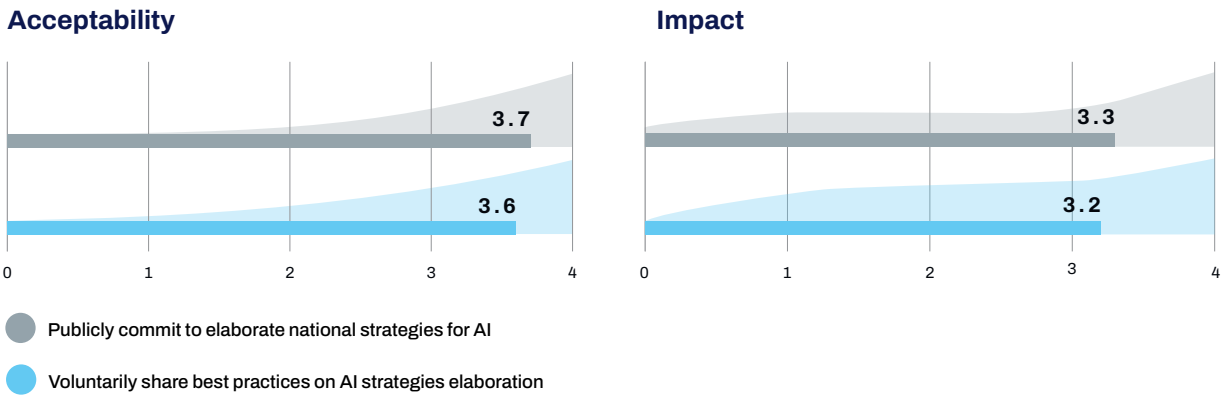
3.2.1. Surveys Results

The figures below present the comparative distribution of scores for each suggested CBM, showing the results for "Acceptability" and "Impact".

The shaded area adjoined to each score line represents the actual distribution of votes, which are stacked along the 1 to 4 values of the scale.

Acceptability		**Impact**	
1	No acceptability	1	No impact
2	Low acceptability	2	Low impact
3	Moderate acceptability	3	Moderate impact
4	High acceptability	4	High impact

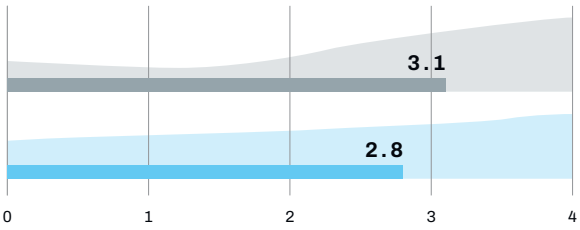
1. Promote the elaboration of national AI strategies (for States which have not yet done so)



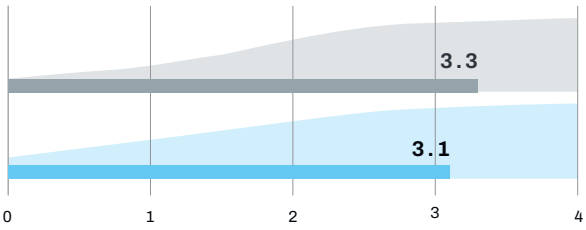
[55] The no – low – moderate – high dimensions were explained as follows—no acceptability/impact: the measure is not acceptable/has no or negligible impact; low acceptability/impact: the measure has very limited levels of acceptability/impact; moderate acceptability/impact: the measure is likely to have medium levels of acceptability/impact and can yield some positive results; high acceptability/impact: the measure will likely have strong acceptability/ strong and positive impact.

2. CBMs for AI safety: Support the convening of technical experts to elaborate common definitions and standards for AI safety

Acceptability



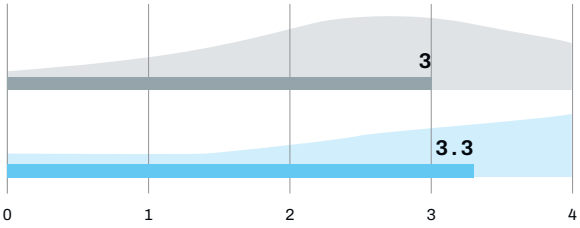
Impact



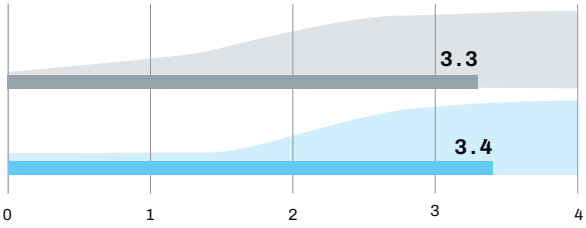
- Promote exchanges at working level between national experts to discuss key terminology and standards related to safety and security of AI systems
- Initiate new international mechanism(s) for exchanges on safety-related issues (e.g., an intergovernmental forum at the international level)

3. Promote legal reviews for AI-enabled weapons systems

Acceptability



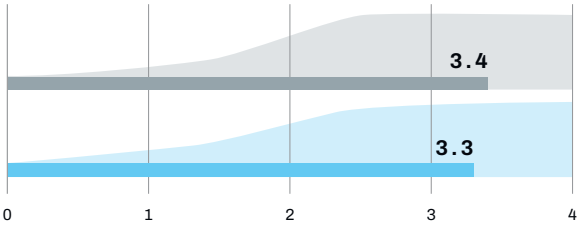
Impact



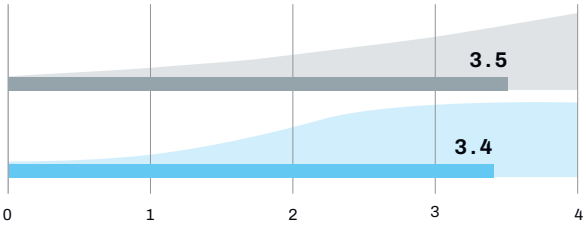
- Publicize high-level information on methodologies and steps for conducting legal reviews; this could include info on the process, department etc.
- Provide funding or sponsor training and capacity-building for experts tasked to review legality of weapons systems

4. Military-to-military dialogues to exchange on doctrines, rules of engagement

Acceptability



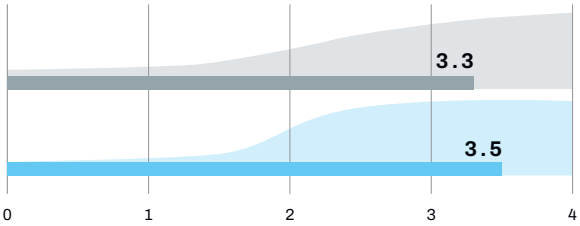
Impact



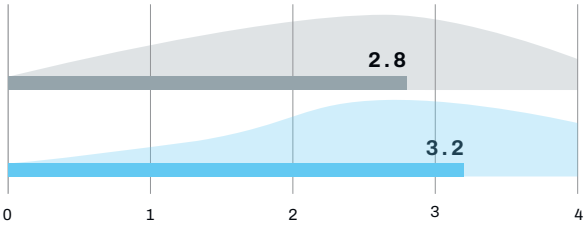
- Exchanges could take place at agreed frequency to discuss scenarios (e.g., involving LAWS etc.) that may result in escalation
- Exchanges could discuss and clarify issues emerging from non-weapons use case, specific risks and mitigation strategies

5. Promote and fund civilian research on AI safety and cybersecurity

Acceptability



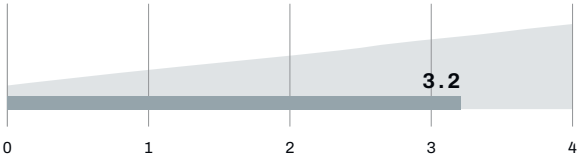
Impact



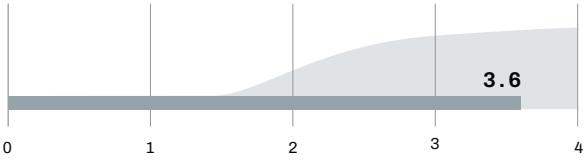
- Unilateral measure: States could commit to bolstering AI safety research in academic centres, and enable civilian leadership of AI safety research
- Multilateral measure: promote Track II exchanges between scientific and academic experts

6. Application or domain-specific CBMs

Acceptability



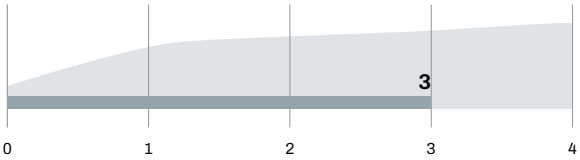
Impact



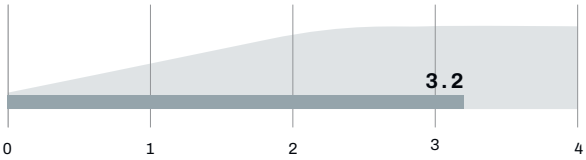
- Identify top priorities and areas of high/unacceptable risks and agree on key principles and approaches (e.g. politically binding documents)

7. (L)AWS-specific CBM

Acceptability



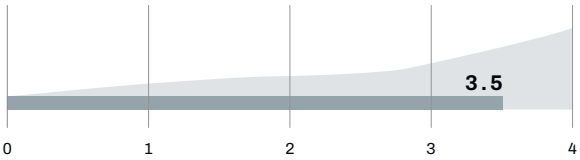
Impact



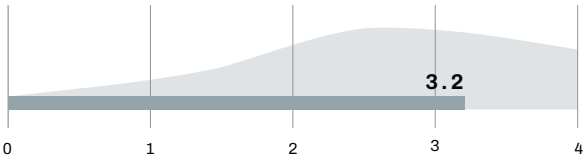
- Code of conduct for autonomous weapons, which could also define ‘rules of the road’, clarify red lines , ways to avoid escalations

8. Share best practices for AI testing and evaluation

Acceptability



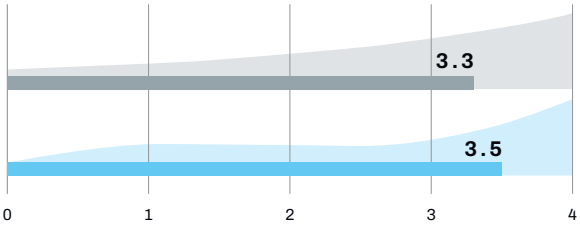
Impact



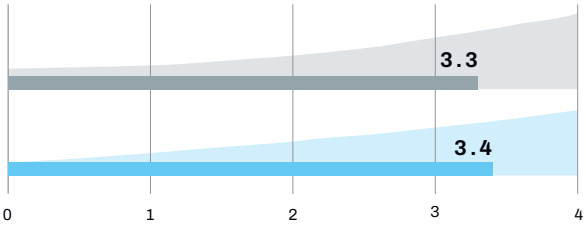
- Publicize, on a voluntary basis, best practices and lessons learned

9. Promote efforts for non-proliferation and prevention of deliberate misuse

Acceptability



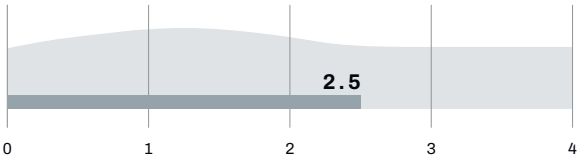
Impact



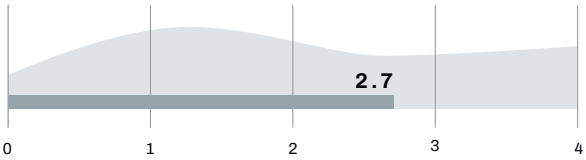
- Unilateral: Strengthen domestic mechanisms for counter-proliferation, which could incorporate risk assessments and mitigation strategies
- Multilateral: Convene experts and promote awareness about AI risks across other processes and disarmament bodies

10. Create a GGE on CBMs for AI

Acceptability



Impact



- Establish a GGE (e.g., at the General Assembly) to promote routinized exchanges on specific themes and deliberate on the elaboration of CBMs

3.2.2. Discussion and Analysis

The distribution of scores for the suggested measures reveals a strong preference for addressing high-risk areas, notably military-to-military exchanges (among the highest ranked measures, both for being politically acceptable and for impact), and non-proliferation and misuse of the technology. Interestingly, the operationalization of the CBM on non-proliferation was deemed even more acceptable and impactful at the multilateral level than within national contexts, which may suggest that many States either consider that sharing knowledge at the international level can help build a foundation for the development of expertise, which can be simultaneously leveraged for the adaptation of national mechanisms, or that many States are confident that their existing domestic mechanisms are sufficiently well-positioned to respond to this challenge and that a more important effort is to streamline these efforts internationally.

The option of developing CBMs over specific applications or domains of use of AI technologies received among the highest scores for impact, and marginally less for acceptability, which may suggest that the measure is acknowledged to be impactful (as it would address concrete concerns in areas of high risk), though slightly less politically acceptable, potentially due to the implied outcome of this measure, which contains a form of commitment (politically binding document). By contrast, the option of military-to-military exchanges, which is also focused on high-risk uses of AI, received a high score both for acceptability and impact, and did not, comparatively, contain an option of further commitment (such as a politically binding document etc.) other than dialogue and exchanges over areas of risks.

The highest scores for 'acceptability', however, were allocated to a unilateral measure, the development of national AI strategies. This may suggest that States consider this measure extremely important, first because many still do not yet have a national strategy on AI,[56] and second, because the political cost of developing a national strategy is lower relative to other suggested CBMs, which may be perceived to have a higher political commitment cost. Of note, the same measure received a slightly lower score for 'impact', which may reinforce the premise that, in the context of building confidence among other States, the net effect of this measure may be considered somewhat less effective even if it is recognized as an important transparency measure and politically highly acceptable.

At the lower ends, among the lowest scores were given to the option of establishing a GGE on CBMs, which was scored low both for acceptability and impact, and to the option of initiating new international mechanisms (e.g., an intergovernmental forum) focused on safety-related issues, which was scored relatively low for acceptability though slightly higher for impact.

It is, however, important to contextualize the apparent lack of support for the initiation of new processes in that some States may be wary of a possible duplication of efforts (or interference with other processes), as well as foresee practical challenges due to lack of capacity, particularly for smaller delegations. The emphasis on capacity-building during the workshop (discussed in the previous section) echoes this concern.

Further, the results which indicate reluctance for creating new institutions need not

[56] The urgent development of national strategies was also one of the recommendations of the Secretary-General in "A New Agenda for Peace". See United Nations, "Our Common Agenda", 28. This is an important unilateral action to consolidate national action plans on responsible development and use of AI.

be misread for a wholesale rejection of the scope of that CBM, which was about AI safety. The other option proposed under this CBM – to promote exchanges over safety concerns between national experts – received a higher score, both for acceptability and impact. Moreover, a related measure, focused on the promotion of civilian research on AI safety was also assessed favorably (though less for impact), and the option to promote Track II exchanges received among the highest scores for acceptability, as well as a high score for impact. This clearly indicates that there is an overall interest to engage in discussions over AI safety, and that several options can be explored but short of creating new international mechanisms or institutions.

These results align with some of the views shared during the workshop, when it was mentioned that most suggested measures are desirable in theory, but that the answers rather reflect assessments of feasibility and real expected benefit in the current context. Other views provided further nuance in that they considered that while important to invest in many discussions, it is also advisable to strike a balance between "too many" and "too few" forums and to avoid "the proliferation of non-proliferation discussions". In addition, it should be mentioned that though the average score for the option of establishing new international mechanisms was low (relative to other measures), it did receive high scores from some of the respondents, and the overall impracticality of this option was not unanimously agreed upon. According to views expressed during the workshop, there can be co-existence between processes, and it may even be inevitable to have parallel and interconnected processes given the complexity of the topic.

Scores for other suggested CBMs were generally situated in the middle, and these included some measures discussed over the years in the GGE on LAWS. The proposal for a CBM for LAWS such as a code of conduct, or on legal reviews for AI-enabled weapons systems were on average assessed to be of moderate acceptability and impact; these scores tie in with polarized views on some of these topics within the GGE. Further, even if the option for a code of conduct on LAWS were in principle considered beneficial (and this measure did receive a slightly higher score for 'impact' compared to 'acceptability'), some States may see promoting it as detrimental to their sustained national position within the GGE on LAWS, where they support efforts for a legally binding instrument on lethal autonomous weapons systems.

On legal reviews, it is noteworthy that the option on supporting capacity-building for legal reviews scored slightly higher, both for acceptability and impact, than the option on transparency. This may suggest a constructive approach to help strengthen national capacities for conducting legal reviews, even as, outwardly, States may be less inclined to share information over their own processes.

Conclusion

The development and proliferation of AI technologies comes with significant transformative potential in the context of international security. As States begin to harness the opportunities afforded by artificial intelligence, the complexity of the risks landscape is being simultaneously acknowledged: the technology can be deliberately misused by malicious actors to create escalatory effects or proliferate new weapons, means and methods of warfare, and it presents numerous risks of malfunctions even when employed by responsible (state) actors.

Confidence-building measures can help address concerns related to the development and use of AI in the context of international security. Confidence-building measures are voluntary, non-legally binding measures that States can take to address either narrow or broad security concerns. CBMs have a long and tested history in the field of arms control and disarmament and can exist either in the absence of, or alongside legally binding instruments and enforceable treaties.

In the field of AI, the discussions on CBMs in the multilateral domain are in an early stage and have thus far been mostly tangential to the subject of lethal autonomous weapons systems in the GGE on LAWS.

The UNIDIR project on CBMs for AI aimed to advance these conversations and to initiate preliminary and substantive points of departure for future deliberations among Member States.

The first phase of this project, which concluded in late 2023, elaborated an AI risks taxonomy in the context of international security. This study helped inform the conceptualization of the subsequent scoping work on CBMs, which aimed to identify first, how Member States evaluate the role and development of CBMs for AI and second, to invite States to provide views on concrete ideas of CBMs in order to assess initial areas of agreement as well as limitations.

This report concludes **the second, and final, phase of the project** and provided a realistic assessment of the role, objectives, and possible pathways for the development of CBMs for AI, drawing directly from perspectives shared by a diverse group of national representatives.

A general conclusion is that States wish to advance conversations about CBMs. This point was unequivocally shared during the open discussions at the workshop, and the surveys too reveal a positive spirit of interest in developing measures to build confidence around the development and use of AI, particularly in areas of high risk. The fundamental challenges, therefore, remain not of intent but rather of substance and degree. The greatest divergences concern how far States are willing to commit and what they consider to be politically feasible in the current context.

Yet, a certain degree of hesitancy at the start of a process may be expected. Striking an optimal balance between what is desirable and what is feasible is a difficult task in any political and multilateral process, and it is critical for States to continue the dialogue and to carve options for action. By way of recommendation, future deliberations may benefit, for example, from deeper dives into specific themes, including through table-top exercises that discuss concrete scenarios and further clarify positions.

The UNIDIR project on CBMs for AI, including the discussion convened as part of the project, is one step in what needs to become an ongoing dialogue. Future UNIDIR work will continue to support these efforts.

Bibliography

Desjardins, Marie-France. “In search of a theory: Developing the concept”. *The Adelphi Papers* Vol 36, Issue 307: Rethinking Confidence-Building Measures (1996). https://www.tandfonline.com/doi/abs/10.1080/05679329608449406.

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.

----- “Indicative timetable of the first 2024 session”. 2024 https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/Indicative_timetable_-_first_GGE_LAWS_session_2024.pdf.

----- “Food for thought paper – Indicative timetable”. 2023 https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/Indicative_timetable_-_first_GGE_LAWS_session_2023.pdf.

------ Statement by Türkiye. 7 March 2023. UN Web TV: https://webtv.un.org/en/asset/k19/k19n8iayzg.

------ Statement by United Kingdom. “Item 5 – Topic 6: Risk mitigation and confidence-building measures”. 9 March 2023. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/UK_Intervention_Item_5_Topic_6_Risk_Mitigation_and_Confidence_Building_Measures.pdf.

----- Draft Proposal “Roadmap Towards New Protocol on Autonomous Weapons Systems”. Submitted by the delegations of Argentina, Costa Rica, Guatemala, Kazakhstan, Nigeria, Panama, Philippines, Sierra Leone, State of Palestine, Uruguay. 2022. https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fdocuments.unoda.org%2Fwp-content%2Fuploads%2F2022%2F05%2F20220311-G10-proposal-legally-binding-instrument.docx&wdOrigin=BROWSELINK.

------ Proposal for an international legal instrument on Lethal Autonomous Weapons Systems (LAWS), Submitted by Pakistan. 8 March 2023. CCW/GGE.1/2023/WP.3/Rev. 1. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.3_REv.1_0.pdf.

------ Working paper by the Bolivarian Republic of Venezuela on behalf of the Non-Aligned Movement and other States Parties to the Convention on Certain Conventional Weapons. 2022. https://documents.unoda.org/wp-content/uploads/2022/08/WP-NAM.pdf.

------ Working Paper of the Russian Federation “Application of International Law to Lethal Autonomous Weapons Systems (LAWS)”. 18 July 2022. https://documents.unoda.org/wp-content/uploads/2022/07/WP-Russian-Federation_EN.pdf.

------ Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. CCW/GGE.1/2019/3. 25 September 2019. https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf.

Mason, Simon J.A. and Matthias Siegfried. "Confidence Building Measures (CBMs) in Peace Processes". In *Managing Peace Processes: Process related questions. A handbook for AU practitioners*. Volume 1, African Union and the Centre for Humanitarian Dialogue. 2013. 57–77. https://peacemediation.ch/wp-content/uploads/2013/07/AU-Handbook_Confidence-Building-Measures-in-Peace-Processes.pdf.

Organization for Security and Co-operation in Europe.

------ OSCE Secretariat. "OSCE Guide on Non-military Confidence-Building Measures (CBMs)". 2012. https://www.osce.org/files/f/documents/6/0/91082.pdf.

------ OSCE Permanent Council. 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.

------ OSCE Permanent Council, 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.

------ Document of the Stockholm Conference on Confidence- and Security-Building Measures and Disarmament in Europe Convened in Accordance with the Relevant Provisions of the Concluding Document of the Madrid Meeting of the Conference on Security and Co-operation in Europe. 19 September 1986. https://www.osce.org/fsc/41238.

------ Vienna Document of the Negotiations on Confidence-and Security-Building Measures. FSC.DOC/1/99. Istanbul, 16 November 1999. https://www.osce.org/files/f/documents/b/2/41276.pdf.

Persi Paoli, Giacomo, Kerstin Vignard, David Danks, Paul Meyer. "Modernizing Arms Control: Exploring responses to the use of AI in military decision-making". UNIDIR, 2020. https://unidir.org/files/2020-08/Modernizing%20Arms%20Control%20Final.pdf.

Puscas, Ioana. "AI and International Security: Understanding the Risks and Paving the Path for Confidence-Building Measures". UNIDIR. 12 October 2023. https://unidir.org/publication/ai-and-international-security-understanding-the-risks-and-paving-the-path-for-confidence-building-measures/.

United Nations. Regional groups of Member States. https://www.un.org/dgacm/en/content/regional-groups.

United Nations General Assembly.

------ Report of the Open-ended Working Group on Security of and in the Use of Information and Communications Technologies 2021–2025. A/78/265. 1 August 2023. https://digitallibrary.un.org/record/4020967?ln=en&v=pdf.

------ Report of the Disarmament Commission for 2023. A/78/42. 27 April 2023. https://documents.un.org/symbol-explorer?s=A/78/42&i=A/78/42_7529841.

------ Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security. A/76/135. 14 July 2021. https://front.un-arm.org/wp-content/uploads/2021/08/A_76_135-2104030E-1.pdf.

------ Report of the Open-ended Working Group on Developments in the Field of Information and Telecommunications in the Context of International Security, A/75/816, 18 March 2021, https://undocs.org/Home/Mobile?FinalSymbol=A%2F75%2F816&Language=E&DeviceType=Desktop&LangRequested=False.

------ "Report of the Disarmament Commission for 2017". A/72/42. https://undocs.org/Home/Mobile?FinalSymbol=A%2F72%2F42.

------ 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.

United Nations Office for Disarmament Affairs.

------ "Securing Our Common Future. An Agenda for Disarmament". 2018. https://front.un-arm.org/wp-content/uploads/2018/06/sg-disarmament-agenda-pubs-page.pdf.

------ "Military Confidence-Building Measures". https://disarmament.unoda.org/convarms/military-cbms/.

------ "Transparency and Confidence Building". https://disarmament.unoda.org/convarms/transparency-cbm/.

------ BWC Confidence Building Measures. https://disarmament.unoda.org/biological-weapons/confidence-building-measures/.

United Nations Secretary-General. "Our Common Agenda. Policy Brief 9. A New Agenda for Peace". July 2023. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf.

@unidir
/unidir
/un_disarmresearch
/unidirgeneva
/unidir

Palais des Nations
1211 Geneva, Switzerland

WWW.UNIDIR.ORG