

PRINCIPLES OF RESPONSIBLE STATE BEHAVIOR WITH RESPECT TO ACTIVITIES WITHIN THE LIFECYCLE OF ARTIFICIAL INTELLIGENCE SYSTEMS IN THE CONTEXT OF INTERNATIONAL PEACE AND SECURITY¹

PRINCÍPIOS DE COMPORTAMENTO ESTATAL RESPONSÁVEL EM RELAÇÃO ÀS ATIVIDADES NO CICLO DE VIDA DE SISTEMAS DE INTELIGÊNCIA ARTIFICIAL NO CONTEXTO DA PAZ E SEGURANÇA INTERNACIONAIS²

NATALIYA MAROZ³

ABSTRACT

Artificial intelligence (AI) systems not only provide a range of opportunities for sustainable development but also may negatively affect international peace and international security. In particular they can be used to target peacekeepers, provoke violence and armed conflicts, magnify the effects of malicious cyber operations and humiliating memory of genocide victims. Taking into account the novelty of risks, inexplicability of decision making of AI, its unpredictability the article attempts to reveal and systematize AI-enabled threats to international peace and security. The article emphasizes that existing international law, in particular, jus cogens norms, principles and norms established in the UN Charter, human rights conventions and other relevant treaties are fully applicable to all types of activities within the lifecycle of AI systems. At the same time their provisions may not comprehensively address all the aspects and specifics of risks produced by such technologies. The article argues that a treaty to regulate safe, secure and responsible use of AI within the UN could establish a comprehensive legal framework on the matter. However, considering continuing uncertainty regarding the risks, opportunities and impacts of AI, the article argues that an

RESUMO

Os sistemas de inteligência artificial (IA) não só oferecem uma série de oportunidades para o desenvolvimento sustentável, como também podem afetar negativamente a paz e a segurança internacionais. Em particular, podem ser utilizados para atacar forças de manutenção da paz, provocar violência e conflitos armados, ampliar os efeitos de operações cibernéticas maliciosas e profanar a memória das vítimas de genocídio. Tendo em conta a novidade dos riscos, a inexplicabilidade dos processos decisórios da IA e a sua imprevisibilidade, o artigo busca revelar e sistematizar as ameaças à paz e à segurança internacionais possibilitadas pela IA. O artigo enfatiza que o direito internacional vigente — em especial as normas de jus cogens, bem como os princípios e normas estabelecidos na Carta da ONU, nas convenções de direitos humanos e em outros tratados relevantes — é plenamente aplicável a todos os tipos de atividades ao longo do ciclo de vida dos sistemas de IA. Ao mesmo tempo, tais disposições podem não tratar de forma abrangente todos os aspectos e especificidades dos riscos produzidos por essas tecnologias. O artigo argumenta que um tratado destinado a regular o uso seguro, protegido e responsável da IA no âmbito da ONU pode-

1 The author gratefully acknowledges the financial support of the FAPEMIG and the institutional support of Federal University of Minas Gerais, which were essential to the preparation of this article.

2 “A autora agradece o apoio financeiro da FAPEMIG e o apoio institucional da Universidade Federal de Minas Gerais, que foram essenciais para a elaboração deste artigo.

3 PhD, Associate Professor Belarusian State University, Faculty of International Relations. Email: nmaroz@bsu.by

idea to conclude such a treaty may be considered quite premature now. For that reason, the article proposes the principles of responsible State behavior with respect to activities within the lifecycle of AI systems in the context of international peace and security as a pragmatic approach to enhance stability and security and support the gradual development of an international normative framework at present.

KEYWORDS: artificial intelligence (AI). cybersecurity. international peace and security. threats to international peace and security. principles of responsible state behavior.

ria estabelecer um arcabouço jurídico abrangente sobre o tema. Contudo, considerando a incerteza persistente em relação aos riscos, oportunidades e impactos da IA, sustenta-se que a conclusão de um tratado dessa natureza pode ser considerada, neste momento, bastante prematura. Por essa razão, o artigo propõe princípios de conduta estatal responsável em relação às atividades desenvolvidas ao longo do ciclo de vida dos sistemas de IA, no contexto da paz e da segurança internacionais, como uma abordagem pragmática para fortalecer a estabilidade e a segurança e apoiar o desenvolvimento gradual de um arcabouço normativo internacional.

PALAVRAS-CHAVE: inteligência artificial (IA). Cibersegurança. paz e segurança internacionais. ameaças à paz e segurança internacionais. princípios de conduta estatal responsável.

I. INTRODUCTION

Artificial intelligence (hereafter – AI) is widely considered not only as a significant technological breakthrough, but also as an enabler for the achievement of all sustainable development goals (United Nations Global Compact, 2025). AI reportedly can contribute between \$10 and \$15 trillion to the global economy by 2030 (United Nations, 2023).

However, the more such technologies are integrated with computer systems employed within critical information, social and economic infrastructure, the more they tend to be vulnerable to different types of threats, thus, negatively impacting national and international security.

AI threats have long been associated with bias, discrimination (Technology and society, 2021) and hallucinations of algorithms (IBM, 2025). Nevertheless, nowadays AI abuses go far beyond that and already affect all actors, including the United Nations, other international organizations, states, private actors, non-governmental organizations, businesses and individuals. In particular, 70% of UN peacekeepers are reportedly affected by mis- and disinformation, which severely hampers their ability to carry out their work. AI-enabled cyber-attacks were also already targeting UN peacekeeping and humanitarian operations (Business and Human Rights Resource Centre, 2025).

AI image fakery creates new challenges for humanitarian organizations (Quelch, 2023), has potential to rewrite the history (Makhortykh & Mann, 2024), escalate violence (Sami), aggravate existing political pressure and interfere into domestic affairs (Redacción Razones de Cuba, 2024) and humiliate the memory of Holocaust victims (Völk & Nguyen, 2025).

Such technologies can easily be accessed by non-state actors, including for terrorist and other criminal purposes (United Nations, 2023). AI is already used for automatization and amplification of the impact of malicious cyber activities (Antropic, 2025). In November 2025 the IT security company Anthropic revealed and disrupted the first reported AI-orchestrated cyber espionage campaign (Anthropic, 2025).

As the UN Secretary-General Antonio Guterres comments “The question is not whether AI will influence international peace and security, but how we will shape that influence” (Mishra, 2025).

The potential threats that AI may pose to global security were analyzed in a number of reports, among which produced by the UN Institute for Disarmament Research (Puscas, 2023), UN Secretary-General’s High-level Advisory Body on AI’s (UN AI Advisory body, 2024); European Union (ENISA, 2020), World Economic Forum (World Economic Forum, 2025); states (Safety and security risks of generative artificial intelligence to 2025 (Annex B), 2025), (Impact of AI on cyber threat from now to 2027), (AI data security, 2025), (International AI Safety Report 2025, 2025) and researchers (Pashentsev, Bazarkina, & Sebekin, 2024), (Horowitz, et al., 2018) etc. However, the risks and threats to international security caused by AI are complex, multifaced and not fully understood. They are constantly evolving, many of which may still be unpredictable (Yampolskiy, 2024). Thus, their consideration in the context of international law requires specific attention.

There hasn’t been concluded any international treaty specifically regulating international obligations of states with respect to AI at universal level. The initiatives to develop international convention on international information security (Convention on International Information Security (concept) (rus), 2011), (Updated Concept of the Convention of the United Nations on Ensuring International Information Security, 2021), (Updated concept for a convention of the United Nations on ensuring international information security : Annex to the letter dated 15 May 2023 addressed to the Secretary-General, A/77/894, 2023) also haven’t received support from a vast number of member states of the UN, unfortunately (Boyko, 2022, p. 17). The Framework Convention on AI and Human Rights, Democracy and the Rule of Law conclude within the Council of Europe in 2024 is the first and currently the only treaty dedicated to the regulation of activities within the lifecycle of AI system (Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024). However, it can’t be considered as a treaty of universal character open for accession to any interested state owing to the requirements stipulated in article 31 (1) of this convention.

Due to exponential use of AI systems in different areas of social life and potential risks they may pose to the enjoyment of human rights there were ad-

opted a substantial number of recommendations and principles to establish soft law regulations for the ethical development and use of AI systems within their life-cycle (Recommendation on the Ethics of AI (UNESCO, 2021), Principles for the Ethical Use of AI in the United Nations System (Chief Executives Board for Coordination, 2022), Principles for responsible stewardship of Trustworthy AI (G 20, 2019), OECD AI Principles (OECD, 2019), Regulatory considerations on AI for health (World Health Organization, 2023), Requirements and framework for AI/ML-based network design optimization in future networks including IMT-2020 (ITU, 2024), ASEAN Responsible AI Roadmap (ASEAN, 2025), Tianjin Declaration of the Shanghai Cooperation Organization (SCO, 2025), G20 AI Principles (G20 AI Principles, 2019) etc. However, they are not binding for states and don't specifically address AI-enabled risks to international peace and security.

Due to “uncertainty around the risks, opportunities and impacts of AI” (Lennon, 2025) and the need to assess “future directions and implications of AI systems and promote scientific understanding” of AI (United Nations Global Compact, 2025, paras 55-56; United Nations General Assembly, 2025, para 1(a)), an initiative conclude a treaty to regulate safe, secure and responsible use of AI within the UN may be considered quite premature at present, whereas voluntary, non-binding norms for responsible State behavior with respect to activities within the lifecycle of AI systems within their lifecycle could contribute to common understanding of the standards of expected conduct in this field, would increase stability and security in the global ICT environment and pave the way for the conclusion of the UN Convention on this matter.

The UN General Assembly plays an active role in addressing the challenges posed by AI and has been specifically considering this issue since 2024 (United Nations Regional Centre for Western Europe, 2025). The UN General Assembly adopted several acts emphasizing the need to inclusively assess and address the potential impact, opportunities and risks of AI systems on sustainable development and the well-being and rights of individuals and emphasizes the importance of international cooperation to promote coordination and compatibility of emerging AI governance frameworks (the Pact for the Future, para 51) (United Nations General Assembly, 2024), recognizing also that the improper or malicious design, development, deployment and use of AI systems, such as without adequate safeguards or in a manner inconsistent with international law, pose risks that could hinder progress towards the achievement of the 2030 Agenda for Sustainable Development and its Sustainable Development Goals (United Nations General Assembly Resolution 78/265, preamble) (United Nations General Assembly, 2024), stressing the commitment to mitigate the risks of AI in full respect of international law, including international human rights law (United Nations General Assembly, para 52), (United Nations Gen-

eral Assembly, 2024). However, neither the Pact for the Future adopted by the UN General Assembly (United Nations General Assembly, 2024) nor the Global Digital Compact annexed to that Pact (United Nations General Assembly, 2024), don't constitute treaties within the meaning of the Vienna Convention on the Law of Treaties. They also don't specifically define how existing international law is applicable to the activities within the lifecycle of AI systems in the context of international peace and security.

Thus, the present article aims to identify AI-enabled threats to international peace and security and suggests a set of principles of responsible behavior with respect to activities within the lifecycle of AI systems in the context of international peace and security. Such principles are elaborated on the basis of existing international law, in particular the UN Charter, human rights law, international humanitarian law and with due regard to norms, rules and principles of responsible state behavior in the use of information and communication technologies (hereafter – ICTs), Pact for the Future and Global Digital Compact.

II. THREATS TO INTERNATIONAL PEACE AND SECURITY THAT MAY BE POTENTIALLY POSED BY AI TECHNOLOGIES

Negative impact of AI on different areas of social life, national and international security is usually assessed through the term of “risk”. There are different approaches to AI risks, which classify them depending on the impact on international security (the technology's main areas of vulnerabilities and potential for misuse, as well as the broader strategic and geopolitical implications of AI in the context of international peace and security (Puscas, 2023)), or more commonly on social well-being (life-circle approach to risk assessment (Kangerter, 2024), society (legal, ethical, governance risks) (Qian, Siau, & Nah, 2024) etc. The evidence-based approach built on tracking and systematization of actual AI incidents was developed within the OECD (OECD, 2025)).

The term risk is mostly used in the context of international environmental law, as well as due diligence obligations in a broader perspective (Townley). As the ICJ held “whether an activity constitutes a risk of significant harm depends on “both the probability or foreseeability of the occurrence of harm and its severity or magnitude” (Advisory Opinion Obligations of States in respect of Climate Change, 2025, para 275). At the same time the term “threat” is mostly applied within the area of law of international security.

The UN AI Advisory Body conceptualizes AI-related risks in relation to vulnerabilities (UN AI Advisory body, 2024, para 17). It distinguishes 4 major areas vulnerable to AI-related risks: individuals, politics and society, economy and environment. Each of these areas includes either targeted sub areas (f.e. rule of law, intellectual property protection etc.), values (human dignity, value or agency etc.) or potential negative influence exerted by the AI (discrimination

and unfair treatment of groups, including based on individual or group traits, such as gender, group isolation and marginalization) (UN AI Advisory body, 2024, p. 27).

A possible negative impact of AI on international peace and security is addressed in the Report of the UN AI Advisory body “Governing AI for humanity” (hereafter – Report), however it is mostly linked to military use of AI systems (UN AI Advisory body, 2024, p. 30). International security is not indicated as a separate area vulnerable to AI-enabled risks within the typology proposed in the Report and has been included in the block “politics and society” (UN AI Advisory body, 2024, p. 27). Risks to international security are considered alongside with national security risks in one block disregarding the fact that these two areas are different in nature and scope (UN AI Advisory body, 2024, p. 27). Such generalized approach may give rise to incorrect interpretation, where national security risks or threats of a particular state may be considered as threats to international peace and security, which could be used to legitimize the adoption of unilateral coercive measures prohibited by international law (Mandate of the Special Rapporteur on the negative impact of unilateral coercive measures on the enjoyment of human rights, 2025, p. 17) or even unilateral military action contrary to the UN Charter (Dörr, 2019). The UN Charter confers exclusive powers to the UN Security Council in the determination of threat to international peace and security, breach of peace and act of aggression and provides it with the exclusive powers to decide how to respond to them, including with military force (United Nations Charter, 1945, article 39, 41-42). Coercive measures can only be taken by States and international organizations in the course of the implementation of resolutions of the UN Security Council adopted under Chapter VII of the UN Charter, or when such measures do not violate their international obligations (retorsions), or their wrongfulness is precluded under international law when the measures constitute counter-measures in full conformity with the rules of the law of international responsibility (Mandate of the Special Rapporteur on the negative impact of unilateral coercive measures on the enjoyment of human rights, 2025, para 17.1). Therefore, possible negative implications caused by AI systems to international peace and security should be addressed separately from national security risks.

Prior to considering AI-enables threats to international peace and security it's necessary to reveal the notion of threat to peace in general, and then more specifically – in the context of AI-enabled activities.

The term “threat” is used in the UN Charter (article 1(1), 2(4), 39, 99 and title of chapter VII) but not defined there (United Nations Charter, 1945). As Wet and Wood comment “it does not have a fixed meaning” and the UN Security Council’s practice shows considerable flexibility on the point” (Wet & Wood, 2022). In its ordinary meaning this term can be defined as “a person

or thing that is regarded as dangerous and likely to inflict harm “, “a strong possibility of something dangerous or unpleasant happening” (Threat). Townly considers a threat “as a form of “known” risk” in his research (Townley). Toledo explains that a threat is “any foreseeable cause of harm to people, property, information or the environment, while risk is the probability of harm or loss to people, property, information and the environment as a result of the threat materialising, taking into account the vulnerability of the elements exposed to such threat” (Toledo, 2025). Verush states that “risk” and “threat” can be considered as different degrees of danger with respect to their provability of occurrence (Verush, 2018, p. 27).

It’s worth noting that the term “risk” is used alongside with the term “threat” in the Best Practices Guide for Chairs and Members of United Nations Sanctions Committees (“new threats and risk actors” (p. vii), “risks to international peace and security” (p. v), “risk-actors such as terrorists, proliferators of WMD, militias and violators of human rights should not gain access to arms, ammunition, mercenaries, dual-use material, commodities of which sale proceeds can be traded against weapons, or coercive employment and human trafficking” (p. 35) etc.) (Best Practices Guide for Chairs and Members of United Nations Sanctions Committees, 2018). However, the Best Practices Guide for Chairs and Members of United Nations Sanctions Committees is not a legally binding act.

Therefore, it’s possible to conclude that the terms “risk” and “threat” have close meaning and reflect different degrees of probability of harm that may occur. However, the term “risk” is not used in the UN Charter. Moreover, only the threat to international peace and security could give rise to the adoption of measures under articles 41-42 by the UN Security Council.

The concept of threat to peace has dramatically changed since the adoption of the UN Charter in 1945. At that time it was mainly associated with international armed conflicts (United Nations Security Council, 2025). The Preamble of the UN Charter sets forth the determination “to save succeeding generations from the scourge of war, which twice in our lifetime has brought untold sorrow to mankind” (United Nations Charter, 1945).

As E.D. Wet and M. Wood comment “the late 20th century and 21st century saw the identification of ‘new’ global threats to international peace and security by the UN Security Council” (Wet & Wood, 2022). The overall perception of threats changed at that time. The Report of the High-level Panel on Threats, Challenges and Change “A more secure world: Our shared responsibility” of 2004 identifies the following new threats to international security: “poverty, infectious disease and environmental degradation; war and violence within States; the spread and possible use of nuclear, radiological, chemical and biological weapons; terrorism; and transnational organized crime” (p.9) (Unit-

ed Nations, 2004). It was also recognized there that “the threats are from non-State actors as well as States, and to human security as well as State security” (United Nations, 2004).

The determination of the existence of any threat to peace an exclusive power of the UN Security Council (article 39 of the UN Charter). The practice of the UN Security Council in the context of application of article 39 of the UN Charter is quite specific. It qualified as a threat to peace inter- or intra-State conflicts or internal conflicts with a regional or sub-regional dimension; as well as terrorist acts, the proliferation of weapons of mass destruction or the proliferation and illicit trafficking of small arms and light weapons (UNSC, n.d.), illegal exploitation of natural resources, where such exploitation was linked to armed conflicts (Wet & Wood, 2022). The UN Security Council determined as a threat to international peace and security widespread and flagrant violations of international humanitarian law, including reports of mass killings, massive, organized and systematic detention and rape of women and the continuance of the practice of “ethnic cleansing”, including for the acquisition and the holding of territory (UN Security Council, 1993), genocide and other serious violations of international humanitarian law (UN Security Council, 1994).

New threats to international peace and security that have been identified by the UN Security Council more recently includes: the unprecedented extent of the Ebola outbreak in Africa, foreign terrorist fighters, violent extremism (UNSC, 2014-2015), the potential for acquisition by non-State actors of chemical weapons, severity of the devastating humanitarian situation, (UNSC, Part VII. Actions with respect to threats to the peace, 2016-2017). During 2018-2023 the Council did not determine new threats to international peace and security (UNSC, Actions with Respect to Threats to the Peace, Breaches of the Peace, and Acts of Aggression, n.d.).

The UN General Assembly identified as threats to international peace and security most serious breaches of international obligations arising out of peremptory norms of international public law (for example, in its resolution 39/19 (Nations, 1984) (apartheid, crimes against humanity), resolution 42/ 95 (violation of the right of peoples to self-determination) (United Nations General Assembly, 1987); resolution 44/101 (continuation of colonialism in all its forms and manifestations-including racism, apartheid, those activities of foreign economic and other interests contrary to the Charter of the UN and the Declaration on the Granting of Independence to Colonial Countries and Peoples, as well as the violations of the right to self-determination and basic human rights of the peoples of colonial Territories and continuous policies and practices to suppress legitimate national liberation movements) (United Nations General Assembly, 1989) (United Nations General Assembly, 1989); resolution 70/146 (recognizes that torture may amount to crimes against humanity or war crimes) (United Na-

tions General Assembly, 2015); resolution 79/1 (violations of the purposes and principles of the Charter) (United Nations General Assembly, 2024). The UN General Assembly also defined certain conduct as a “danger” to international peace and security (for instance, resolution 40/150 (arms race, particularly in nuclear armaments and military expenditures) (United Nations General Assembly, 1985).

Negative AI implications haven’t been qualified as a threat to international peace and security by the UN Security Council yet. However, the topic was discussed at its formal meetings and informal meetings of the members of the UN Security Council (UN Security Council, 2024).

In particular, during the Arria-formula Meeting on “AI: Its Impact on Hate Speech, Disinformation and Misinformation” the UN Security Council acknowledged that AI can “compound threats and intensify conflicts by spreading hate speech, undermining the integrity of public information, including the interpretation of international humanitarian law, amplifying bias and exacerbating human rights abuses, contributing to increased polarization and instability on a vast scale” (Security Council Report, 2023). In its resolution 2686 (2023) it admitted that “hate speech, racism, racial discrimination, xenophobia, related forms of intolerance, gender discrimination, and acts of extremism can contribute to driving the outbreak, escalation and recurrence of conflict” (United Nations Security Council, 2023).

Moreover, “malicious actors could exploit AI for cyberattacks, disinformation, misinformation or extremist rhetoric, to interfere in elections or contribute to increased intercommunal tension” (Security Council Report, 2023). During the Arria-formula meeting “Harnessing safe, inclusive, trustworthy AI for the maintenance of international peace and security” it was recognized that AI can “undermine trust in peacekeepers, propagate misinformation and hate speech, and facilitate malicious cyberattacks” (Security Council Report, 2025).

It’s also widely accepted that existing and emerging threats from uses of ICTs, by States and non-State actors may jeopardize international peace and security (United Nations, 2015).

The UN General Assembly has repeatedly emphasized that ICTs can be used for purposes incompatible with the objectives of ensuring international stability and security and can negatively impact the security of States (resolutions 53/70 of 4 December 1998, 54/49 of 1 December 1999, 55/28 of 20 November 2000, 56/19 of 29 November 2001) in both the military and civilian spheres (resolutions 56/19 of 2001, 57/53 of 22 November 2002, 58/32 of 8 December 2003, 59/61 of 3 December 2004, 60/45 of 8 December 2005, 61/54 of 6 December 2006, 62/17 of 5 December 2007, 63/37 of 2 December 2008, 64/25 of 2 December 2009, 65/41 of 8 December 2010, 66/24 of 2 December 2011, 67/27 of 3 December 2012, 68/243 of 27 December 2013, 69/28 of 2 De-

cember 2014, 70/237 of 23 December 2015, 71/28 of 5 December 2016, 73/27 of 5 December 2018, 74/29 of 12 December 2019, 76/19 of 6 December 2021, 78/237 from December 22, 2023).

The UN General-Secretary mentioned the following AI threats, including in the context of international security: military application of AI and AI arms race; interaction between AI and nuclear weapons, biotechnology, neurotechnology and robotics; malicious use of AI systems for terrorist, criminal or State purposes; AI-enabled cyberattacks targeting critical infrastructure, peacekeeping and humanitarian operations; disinformation and hate speech; deep fakes; malfunctioning of AI systems; the environmental impact of AI (United Nations, 2023).

The UN General Assembly resolution specifically distinguished “improper or malicious design, development, deployment and use of AI systems, such as without adequate safeguards or in a manner inconsistent with international law” as a risk that could hinder progress towards the achievement of the 2030 Agenda for Sustainable Development and its Sustainable Development Goals and undermine sustainable development in its three dimensions – economic, social and environmental; widen digital divides between and within countries; reinforce structural inequalities and biases; lead to discrimination; undermine information integrity and access to information; undercut the protection, promotion and enjoyment of human rights (para 4) (UN General Assembly, 2024).

Thus, based on the aforementioned it's possible to distinguish the following AI-enabled threats to international peace and security

- use of AI for massive targeted cyber-attacks against critical infrastructure of states, scale and effects of which commensurate to armed attack,
- propaganda for war, misinformation, disinformation, manipulation of data, including with the use of deep fakes provoking escalation, leading to international armed conflict as well as to armed conflict of non-international character,
- misinformation, disinformation, manipulation of data, including with the use of deep fakes aimed at intimidation, escalation of military presence and all other forms of intervention or interference, direct or indirect, overt or covert, threatening the sovereignty and political independence of States, including with the aim of overthrowing their Governments as well as aimed at deprivation of peoples of their right to self-determination,
- AI arms race,

- integration of AI and nuclear weapons or any prohibited types of weapons of mass destruction,
- integration between AI and biotechnology, neurotechnology and robotics for malicious purposes or without exercising due diligence with respect to the negative effects they may provoke,
- use of AI for targeting peacekeeping and humanitarian operations both in peace time and during armed conflict,
- use of AI operation for perfidy, for targeting impartial humanitarian organizations, civilian objects, including medical facilities, works and installations containing dangerous forces, objects indispensable for the civilian populations during armed conflicts in violation of international humanitarian law,
- use of AI for widespread and systematic human rights violations, direct and public incitement to commit genocide,
- use of AI for malicious purposes by terrorist and transnational criminal groups;
- availability of autonomous lethal weapons of terrorists and transnational organized groups,
- use of AI agents for terrorist or criminal purposes, including for “conscription” of other AI agents or use of AI technologies for malicious purposes,
- undermining integrity of international law, by misinformation, disinformation, manipulation of data, including with the use of deep fakes; falsifying history about the suffering of victims of racism, racial discrimination, xenophobia and related intolerance and forms of discrimination in the historical perspective, in particular regarding commemoration of the victims of slavery and the transatlantic slave trade; for revisionism and attempts to falsify history, including for the purpose of glorification of Nazism or denial of crimes established by the International Military Tribunal in Nuremberg.

III. RESPONSIBLE STATE BEHAVIOR WITH RESPECT TO ACTIVITIES WITHIN THE LIFECYCLE OF AI SYSTEMS IN THE CONTEXT OF INTERNATIONAL PEACE AND SECURITY

Despite the fact AI-enabled risks and threats are new, the existing international law, including the UN Charter is fully applicable to respond to them (UN Security Council, 2024, p. 15), (United Nations Global Compact, 2025),

though, may not address all peculiar challenges arising out of the use of AI systems.

Peremptory norms of international law don't permit any derogation from them (Vienna Convention on the Law of Treaties, 1969, article 53), (International Law Commission, 2022, Conclusion 4) and, therefore, any AI-enabled activities constituting racial discrimination, breach of the right to self-determination or constituting violation of prohibition of genocide, apartheid, aggression, crimes against humanity, slavery, torture, basic rules of international humanitarian law with a reference to newly emerged character of AI systems or their unpredictability, lack of explainability or transparency in decision making can't be invoked as justification for its failure to fulfil obligation under peremptory norm of general international law.

International obligations of states with respect to the use of AI derive from paras 2,3 of Article 1 (principle of equal rights and self-determination of peoples; principle of cooperation; respect for human rights and for fundamental freedoms), paras 1, 4, 7 of Article 2 of the UN Charter (principle of sovereign equality; the prohibition of the threat or use of force against the sovereignty, territorial integrity or political independence of any State; the principle of non-intervention in the domestic affairs) (United Nations Charter, 1945), Article 20 of the International Covenant on Civil and Political Rights (the prohibition of propaganda for war, advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence) (International Covenant on Civil and Political Rights, 1966), Articles 51-52 of the Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field of 1949 (Geneva Convention for the amelioration of the condition of the wounded and sick in armed forces in the field, 1949), Articles 40, 51 of the Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I) (the prohibition of acts or threats of violence the primary purpose of which is to spread terror among the civilian population, threats that there will be no survivors, threats to commit acts provided for in Article 75 of Protocol I) (Protocol additional to the Geneva Conventions of 12 August 1949, and relating to the protection of victims of international armed conflicts (Protocol I), 1977); Article III of the 1948 Convention on the Prevention and Punishment of the Crime of Genocide (the prohibition of direct and public incitement to commit genocide) (Convention on the Prevention and Punishment of the Crime of Genocide, 1948), Article III of the 1973 International Convention on the Suppression and Punishment of the Crime of Apartheid (the prohibition of incitement to commit apartheid) (International Convention on the Suppression and Punishment of the Crime of Apartheid, 1973), paragraph 2 of Article 27, paragraph 2 of Article 30 of the Vienna Convention on Diplomatic Relations

(violation of the inviolability of diplomatic correspondence) (Vienna Convention on Diplomatic Relations, 1961), paragraph 2 of Article 35 of the Vienna Convention on Consular Relations (violation of the inviolability of consular correspondence) (Vienna Convention on Consular Relations, 1963), Article 109 of the UN Convention on the Law of the Sea (unauthorized broadcasting from the high seas) (United Nations Convention on the Law of the Sea, 1982), Article 3 bis (a) of the Convention on International Civil Aviation (prohibition of the use of weapons against civilian aircrafts in flight, Article 38 (obligation to safeguard communication channels and installations within their jurisdiction; take such steps as may be necessary to ensure the establishment, under the best technical conditions, of the channels and installations necessary to carry on the rapid and uninterrupted exchange of international telecommunications) (Convention on International Civil Aviation, 1944), as well as in accordance with Articles 34, 35, 37, 39, 45 (1) of the Constitution of the International Telecommunication Union (Constitution and Convention of the International Telecommunication Union, 1992), etc.

Moreover, rules of international customary international law (Wood & Sender, 2024, p. 1), such as the right to diplomatic protection (International Law Commission, 2006, p. 50), the right to self-defense (Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America) (Merits), 1986, para 34), the principle of due diligence (Corfu Channel (United Kingdom of Great Britain and Northern Ireland v. Albania), 1949, para 22), (Ridings, 2024), obligations of states in respect of climate change (Advisory Opinion Obligations of States in respect of Climate Change, 2025, pp. 132-142) etc. are also applicable to the area discussed.

The International Convention Concerning the Use of Broadcasting in the Cause of Peace is a unique international instrument in terms of providing specific international obligations of State Parties to prevent “broadcasting from being used in a manner prejudicial to good international understanding” (International Convention concerning the Use of Broadcasting in the Cause of Peace, 1936). It prohibits broadcasting inciting the population of any territory to acts incompatible with the internal order or the security of a territory of a High Contracting Party (Article 2); incitement to war (Article 3) (International Convention concerning the Use of Broadcasting in the Cause of Peace, 1936). The Convention doesn’t make any difference between broadcasting conducted by States or private companies within their jurisdiction or effective control. It obligates its State Parties to stop “any transmission likely to harm good international understanding by statements the incorrectness of which is or ought to be known to the persons responsible for the broadcast” (Article 3); to ensure “especially in time of crisis, that stations within their respective territories shall broadcast information concerning international relations the accuracy of which

shall have been verified - and that by all means within their power - by the persons responsible for broadcasting the information” (Article 4) (International Convention concerning the Use of Broadcasting in the Cause of Peace, 1936). This treaty may be applicable for tackling disinformation and misinformation with the use of AI in broadcasting to prevent any political tensions from occurrence or aggravation of existing ones. However, the number of State Parties to the International Convention Concerning the Use of Broadcasting in the Cause of Peace is relatively small (only 11 states) (International Convention concerning the Use of Broadcasting in the Cause of Peace. Participants, 2025).

The UN Convention against cybercrime is partly applicable to the matter as far as it concerns cybercrimes indicated in the convention committed with the use of AI systems (articles 7-16) (United Nations Convention against Cybercrime; Strengthening International Cooperation for Combating Certain Crimes Committed by Means of Information and Communications Technology Systems and for the Sharing of Evidence in Electronic Form of Serious Crimes, 2024), nevertheless, doesn't specifically address cybercrimes committed by AI agents. At the same time such behavior could be prohibited in the additional protocol to this convention to be drafted during the upcoming several years (United Nations General Assembly, 2024, para. 5).

The first and to date the only international treaty on AI is the Council of Europe Framework Convention on AI and human rights, democracy and the rule of law (hereafter – The Framework Convention), which was signed by 17 states and the European Union (as for the 16th of December 2025)(Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024) The Framework Convention hasn't come into force yet.

Under article 1(1) of the Framework Convention its goal “to ensure that activities within the lifecycle of AI systems are fully consistent with human rights, democracy and the rule of law” (Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024) The Framework Convention provides the notion of AI systems (article 2), principles related to activities within the lifecycle of AI systems (Chapter III), remedies for violations of human rights resulting from the activities within the lifecycle of AI systems (Chapter IV), assessment and mitigation of risks and adverse impacts (Chapter V), follow-up mechanism and rules for international cooperation between state parties (Chapter VI) (Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024)

The Framework Convention is applicable to the activities within the lifecycle of AI systems that have the potential to interfere with human rights, democracy and the rule of law carried out by States, private actors acting on their behalf and private actors (article 3(1)(a,b). Moreover, States Parties may declare to be bound by the relevant Convention provisions with respect to activities

of private actors or, by taking other appropriate measures to comply with the treaty's provisions and "may not derogate from or limit the application of its international obligations undertaken to protect human rights, democracy and the rule of law (article 3(1)(b)) (Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024)

The Framework Convention also aims to ensure wide cooperation in the realisation of its purpose, including not only between State Parties, but also with other relevant stakeholders and States that are not Parties to the Convention. In particular, the Framework Convention encourages States Parties to "assist States that are not Parties to this Convention in acting consistently with the terms of this Convention and becoming a Party to it" (article 25(1)), to involve, as appropriate, relevant stakeholders and States that are not Parties to the Convention in exchanges of Information relevant and useful information between themselves concerning aspects related to AI which may have significant positive or negative effects on the enjoyment of human rights, the functioning of democracy and the observance of the rule of law, including risks and effects that have arisen in research contexts and in relation to the private sector (article 25 (2)) (Council of Europe Framework Convention on Artificial Intelligence and Human, 2024). However, as has been mentioned before the accession to this Convention is subject to several requirements set forth in article 31 (Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 2024)

There are also regional treaties concluded with the aim to enhance international cooperation in the field of cybersecurity (Agreement on Cooperation in Ensuring International Information Security between the Member States of the Shanghai Cooperation Organization, Agreement on cooperation between the member states of the Commonwealth of Independent States in the field of ensuring information security, Agreement on cooperation between the member states of the Collective Security Treaty Organization in the field of ensuring information security, Agreement on Ensuring Information Security within the Framework of General Customs Processes in the Member States of the Eurasian Economic Community, African Union Convention on Cyber Security and Personal Data Protection). They can partially cover AI-related issues as far as it concerns countering threats to information security or cybercrime. However, none of these treaties address AI-enabled threats specifically.

The provisions on international cooperation on cybersecurity issues, applicable to AI systems, are also contained in some free trade agreements (for example, Article 14.16 (b) of the Comprehensive and Progressive Agreement for Trans-Pacific Partnership (using existing collaboration mechanisms to cooperate to identify and mitigate malicious intrusions or dissemination of malicious code that affect the electronic networks of the Parties (Comprehensive

and Progressive Agreement for Trans-Pacific Partnership, 2018), Article 12.13 (b) of the Regional Comprehensive Economic Partnership Agreement (using existing collaboration mechanisms to cooperate on matters related to cyber security) (Regional Comprehensive Economic Partnership, 2020), article 5.1 (2) (b) of the Digital Economy Partnership Agreement (using existing collaboration mechanisms to cooperate to identify and mitigate malicious intrusions or dissemination of malicious code that affect the electronic networks of the Parties) (The Digital Economy Partnership Agreement (DEPA), 2020) etc.).

Certain aspects of responsible use of AI are regulated by legally binding acts of the EU. In particular, the Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on AI (hereafter – AI Act) introduced a risk-based AI classification system, establishing obligations for providers and users depending on the level of risk of AI risk qualification (Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence , 2024). The Regulation (EU) 2024/1689 prohibits certain unacceptable AI practices such as manipulative, exploitative and social control practices (para 28 of Preamble); biometric categorisation systems that are based on natural persons' biometric data (para 30 of Preamble), social scoring of natural persons by public or private actors (para 31 of Preamble), the use of AI systems for 'real-time' remote biometric identification of natural persons in publicly accessible spaces for the purpose of law enforcement except in exhaustively listed and narrowly defined situations, where the use is strictly necessary to achieve a substantial public interest, the importance of which outweighs the risks (paras 32-33 of Preamble), the use of AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or closed-circuit television footage (para 43 of Preamble), the use of AI systems intended to be used to detect the emotional state of individuals in situations related to the workplace and education (para 44 of Preamble) (Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence , 2024).

Soft law regulation of responsible state behavior in the use of ICTs could be also partially applied to tackling AI-enabled threats to international peace and security.

The issues of responsible state behavior in the use of ICTs has been discussed within the UN since 1998, when the Russian Federation introduced a draft resolution on the subject in the First Committee of the UN General Assembly (Office for Disarmament Affairs , 2025). Since 2004, six Groups of Governmental Experts (GGEs) have assessed how the use of ICTs may threaten international security and considered appropriate ways to respond (Office for Disarmament Affairs , 2025). The reports contain a set of norms, rules and prin-

ciples of responsible state behavior in the use of ICT (Developments in the field of information and telecommunications in the context of international security, 2023) that alongside with the report by the Group of Governmental Experts on Advancing responsible State behavior in cyberspace in the context of international security in 2021 (Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, 2021) and the reports presented by the UN Open Ended Working Group on development in the field of information and telecommunications in the context of international security in 2022 (Open-ended working group on security of and in the use of information and communications technologies 2021–2025, 2022), 2023 (Open-ended working group on security of and in the use of information and communications technologies 2021–2025, 2023) and 2024 (Open-ended working group on security of and in the use of information and communications technologies 2021–2025, 2024) can also be applicable to activities within several stages of lifecycle of AI (in particular, to its development and use), but, however don't specifically address the issues of disinformation and misinformation, falsifications and manipulations of data for malicious purposes, use of autonomous AI agent, interaction between AI and weapons of mass destruction, between AI and biotechnology, neurotechnology and robotics for malicious purposes; AI-enabled attacks against peacekeeping and humanitarian operations; AI arms race.

Therefore, based on the international instruments and non-binding norms, rules and principles analyzed above it's possible to suggest the following Principles of responsible behavior with respect to activities within the lifecycle of AI systems in the context of international peace and international security:

Principle 1. The UN plays a leading role in AI governance, defining rules for responsible state behavior with respect to activities within the lifecycle of AI systems and enhancing confidence building, promoting capacity building for ensuring equal, fair, non-discriminatory and inclusive access of all states to AI technology for sustainable development.

Principle 2. Any activities within the lifecycle of AI systems shall not be carried out contrary to the obligations under the UN Charter, especially in the way it endangers international peace and security. States should seek to prevent the proliferation of malicious AI models and tools.

States should not conduct or knowingly support AI-enabled activity contrary to its obligations under international law that constitute propaganda for war, misinformation, disinformation, manipulation of data, including with the use of deep fakes aimed at intimidation, escalation of military presence and all other forms of intervention or interference, direct or indirect, overt or covert; threatening the sovereignty and political independence of States, including with the aim of overthrowing their Governments; as well as aimed at deprivation of

peoples of their right to self-determination; provoking escalation or leading to armed conflict.

States shall refrain from any violation of international humanitarian law with respect to any activities within the lifecycle of AI systems, including by conducting AI-enabled operations against peacekeeping and humanitarian operations; targeting civilian objects, including medical facilities, impartial humanitarian organizations, works and installations containing dangerous forces objects indispensable for the civilian populations during armed conflicts or peacetime. States shall refrain from integrating AI systems and nuclear weapons or any prohibited types of weapons of mass destruction.

Principle 3. All states are obliged to respect, promote and protect all human rights within the lifecycle of AI systems. States should protect against human rights abuses caused by any activities within the lifecycle of AI systems by third parties, including businesses under their jurisdiction or effective control.

Principle 4. All international disputes arising out from AI-enabled activities shall be settled by peaceful means in such a manner that international peace and security are not endangered. In case of incidents caused by any activities with the use AI systems States should consider all relevant information, including the larger context of the event, the challenges of attribution in the ICT environment and the nature and extent of the consequences. Any means of pressure taken by States in the course of counter-measures in response to international wrongful act committed within the lifecycle of AI systems shall be performed in full compliance with norms of law of international responsibility.

States shall not apply unilateral coercive measures affecting any peaceful activities within the lifecycle of AI systems by other states.

Principle 5. States should not knowingly allow their territory to be used for internationally wrongful acts using AI systems

States should not conduct or knowingly support any AI-enabled activities contrary to its obligations under international law that intentionally damages critical infrastructure or otherwise impairs the use and operation of critical infrastructure to provide services to the public.

Principle 6. States should take all necessary measures to ensure safe, secure, non-discriminative and trustworthy AI systems under their jurisdiction or effective control, including by maintaining measures for the identification, assessment, prevention and mitigation of risks posed by AI systems by considering actual and potential impacts on human rights, international peace and security and sustainable development.

AI systems with unacceptable risks for the enjoyment of human rights, international peace and security shall be prohibited. AI systems with a high risk to international peace and security shall be prohibited. Human oversight of AI systems with a high risk to the enjoyment of human rights shall be guaranteed.

States should take appropriate measures to protect their critical infrastructure from AI-enabled threats.

Principle 7. States shall refrain from any AI-enabled activities undermining integrity of international law by misinformation, disinformation, manipulation of data, including with the use of deep fakes; falsifying history about the suffering of victims of racism, racial discrimination, xenophobia and related intolerance and forms of discrimination in the historical perspective, in particular regarding commemoration of the victims of slavery and the transatlantic slave trade; for revisionism and attempts to falsify history, including for the purpose of glorification of Nazism or denial of crimes established by the International Military Tribunal in Nuremberg.

Principle 8. States shall prevent significant harm to the environment by exercising due diligence and using all means at their disposal to prevent in their activities within the lifecycle of AI carried out within their jurisdiction or control from causing said harm.

Principle 9. States should respond to appropriate requests for assistance by another State whose critical infrastructure is subject to any harm caused by the activities within lifecycle of AI systems. States should also respond to appropriate requests to mitigate such harm aimed at the critical infrastructure of another State emanating from their territory, taking into account due regard for sovereignty.

Principle 10. States should encourage responsible reporting of AI vulnerabilities and share associated information on available remedies to such vulnerabilities to limit and possibly eliminate potential threats to the enjoyment of human rights, international peace and security.

Principle 11. States should cooperate in preventing and combating malicious AI-enabled activities, especially carried out by terrorist groups, transnational criminal organized groups. States respond to appropriate requests for assistance by another State in such cases.

Principle 12. States are responsible for violations of international law, including their human rights obligations, arising out from the activities within the lifecycle of AI systems.

CONCLUSION

AI systems may be used in a manner inconsistent with international law, which can endanger national and international security.

In view of uncertainty surrounding the risks, opportunities and impacts of AI, and the need to assess future directions and implications of AI systems, any efforts to conclude a treaty of universal character governing safe, secure and responsible use of AI by states are premature now. At the same time voluntary, non-binding principles of responsible State behavior with respect to activities

within the lifecycle of AI systems may enhance stability and security in the global information and communications technology environment, and facilitate, at a later stage, the conclusion of the United Nations convention in this area.

The Council of Europe Framework Convention on AI and human rights, democracy and the rule of law can't serve as a treaty of universal character since it's not open for all interested states and sets a series of conditions to be fulfilled for the accession to this convention by non-member states of the Council of Europe.

The principles of responsible state behavior with respect to activities within the lifecycle of AI systems in the context of international peace and security proposed in this article have been elaborated on the basis of principles of international law enshrined in the UN Charter (sovereign equality of states, prohibition of intervention into domestic affairs, non use of force or the threat to use force etc.), jus cogens norms, existing legal instruments applicable to this issue, soft law, including resolutions of the UN General Assembly, the principles for ethical development and use of AI developed by international organizations; norms, rules and principles of responsible state behavior in cyberspace etc. The principles are developed within a threat-based approach. In particular, a list of potential AI-enabled threats to international peace and security has been identified, which is grounded on the UN Security Council practice on the determination of the existence of a threat to the peace, applicable resolutions of the General Assembly, reports on AI-enabled risks, statements of the Secretary-General and other relevant sources.

REFERENCES

Access to electricity (% of the population). (2025). Retrieved from World Bank Group data: <https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS>

Advisory Opinion Obligations of States in respect of Climate Change (International Court of Justice July 23, 2025).

AI data security. (2025, May 23). Retrieved from Australian Signals Directorate : <https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/ai-data-security>

Anthropic. (2025, November). Disrupting the first reported AI-orchestrated cyber espionage campaign, Full report. Retrieved from <https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf>

Antropic. (2025, August 27). Retrieved from Detecting and countering misuse of AI: August 2025: <https://www.anthropic.com/news/detecting-countering-misuse-aug-2025>

Arrest Warrant of 11 April 2000 (Democratic Republic of the Congo v. Belgium) (International Court of Justice 14 <https://www.icj-cij.org/sites/default/files/case-related/121/121-20020214-JUD-01-00-EN.pdf>, 2002).

ASEAN. (2025). ASEAN Responsible AI Roadmap. Retrieved from <https://asean.org/wp-content/uploads/2025/02/ASEAN-Responsible-AI-Roadmap-Final.docx.pdf>

Best Practices Guide for Chairs and Members of United Nations Sanctions Committees. (2018). Retrieved from Sanctions Research Platform: <https://sanctionsplatform.ohchr.org/record/4358>

Boyko, S. (2022). Political and Legal Framework of the International Information Security System: Russian Approaches and Initiatives. *Russian Journal of World Politics and Law of Nations*, 1(1-2), 4-22.

Business and Human Rights Resource Centre. (2025). UN Secretary General says malicious use of AI could cause horrific damage & calls for new UN body to tackle the risks. Retrieved from <https://www.business-humanrights.org/en/latest-news/un-secretary-general-says-malicious-use-of-ai-could-cause-horrific-damage-call-for-new-un-body-to-tackle-the-risks/>

Chief Executives Board for Coordination. (2022). Principles for the Ethical Use of Artificial Intelligence in the United Nations System (Advanced unedited version). Retrieved from United Nations: https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf

Comprehensive and Progressive Agreement for Trans-Pacific Partnership. (2018, March 8). Retrieved from United Nations Treaty Collection: https://treaties.un.org/Pages/showDetails.aspx?objid=080000028056a333&clang=_en

Constitution and Convention of the International Telecommunication Union. (1992, December 22). Retrieved from United Nations Treaty Collection : https://treaties.un.org/Pages/showDetails.aspx?objid=08000002800b0730&clang=_en

Convention on International Civil Aviation. (1944, December 7). Retrieved from United Nations Treaty Collection : https://treaties.un.org/Pages/showDetails.aspx?objid=0800000280163d69&clang=_en

Convention on International Information Security (concept) (rus). (2011). Retrieved from ПИР-Центр: <https://pircenter.org/editions/konvencija-ob-obespechenii-mezhdunarodnoj-informacionnoj-bezopasnosti-koncepcija-rf/>

Convention on the Prevention and Punishment of the Crime of Genocide. (1948, December 9). Retrieved from United Nations Treaty Collection: https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-1&chapter=4

Corfu Channel (United Kingdom of Great Britain and Northern Ireland v. Albania) (International Court of Justice April 9, 1949).

~~Ministry for Foreign Affairs of Japan~~ <https://www.mofa.go.jp/mofaj/files/100792866.pdf> Developments in the field of information and telecommunications in the context of international security. (2023). Retrieved from United Nations Office for Disarmament Affairs: <https://disarmament.unoda.org/en/our-work/emerging-challenges/developments-field-information-and-telecommunications-context>

Dörr, O. (2019). Use of Force, Prohibition of. Retrieved from Max Planck Encyclopedias of International Law : <https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e427>

ENISA. (2020, December). AI Cybersecurity Challenges. Threat Landscape for Artificial Intelligence. Retrieved from <https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Artificial%20Intelligence%20Cybersecurity%20Challenges.pdf>

Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. (2024, September 5). Retrieved from : Ministry for Foreign Affairs of Japan <https://www.mofa.go.jp/mofaj/files/100792866.pdf>

G 20. (2019). Principles for responsible stewardship of Trustworthy AI. Retrieved from Ministry for Foreign Affairs of Japan: https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf

G20 AI Principles. (2019). Retrieved from Ministry of Foreign Affairs of Japan: https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf

Geneva Convention for the amelioration of the condition of the wounded and sick in armed forces in the field. (1949, August 12). Retrieved from United Nations Treaty Collection: <https://treaties.un.org/pages/showDetails.aspx?objid=080000028015847c>

Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security. (2021). Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security, A/76/135. Retrieved from United Nations: <https://docs.un.org/en/A/76/135>

Horowitz, M., Scharre, P., Allen, G., Frederick, K., Cho, A., & Saravalle, E. (2018, July). Artificial Intelligence and International Security. Retrieved from https://s3.us-east-1.amazonaws.com/files.cnas.org/hero/documents/CNAS-AI-and-International-Security-July-2018_Final.pdf

IBM. (2025). Retrieved from What are AI hallucinations?: <https://www.ibm.com/think/topics/ai-hallucinations>

Impact of AI on cyber threat from now to 2027. (n.d.). Retrieved from National Cyber Security Centre: <https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027>

Individuals using the Internet (% of the population). (2025). Retrieved from World Bank Group data: <https://data.worldbank.org/indicator/IT.NET.USER.ZS>

International AI Safety Report 2025. (2025). Retrieved from <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025>

International code of conduct for information security: annex to the letter dated 12 September 2011 to the United Nations addressed to the Secretary-General, A/66/359. (2011, September 12). Retrieved from United Nations: <https://docs.un.org/en/a/66/359>

International Convention concerning the Use of Broadcasting in the Cause of Peace. (1936, September 23). Retrieved from United Nations Treaty Collection : https://treaties.un.org/Pages/showDetails.aspx?objid=0800000280046246&clang=_en

International Convention concerning the Use of Broadcasting in the Cause of Peace. Participants. (2025). Retrieved from United Nations: https://treaties.un.org/Pages/showDetails.aspx?objid=0800000280046246&clang=_en

International Convention on the Suppression and Punishment of the Crime of Apartheid. (1973, November 30). Retrieved from United Nations Treaty Collection: https://treaties.un.org/pages/viewdetails.aspx?src=treaty&mtdsg_no=iv-7&chapter=4&clang=_en

International Covenant on Civil and Political Rights. (1966, December 16). Retrieved from United Nations Treaty Collection : https://treaties.un.org/PAGES/ViewDetails.aspx?chapter=4&clang=_en&mtdsg_no=IV-4&src=TREATY

International Law Commission . (2022). Draft conclusions on identification and legal consequences of peremptory norms of general international law (jus cogens). Yearbook of the International Law Commission, II, Part Two.

International Law Commission . (2006). Draft Articles on Diplomatic Protection with commentaries . Retrieved from https://legal.un.org/ilc/texts/instruments/english/commentaries/9_8_2006.pdf

ITU. (2024). Recommendation ITU-T Y.3142. Retrieved from https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-Y.3142-202404-I!!PDF-E&type=items

Kangeter, L. (2024, June). A Lifecycle Approach to AI Risk Reduction. Retrieved from Institute for security and technology : <https://securityandtechnology.org/wp-content/uploads/2024/06/A-Lifecycle-Approach-to-AI-Risk-Reduction.pdf>

Lennon, C. (2025, September 25). UN moves to close dangerous void in AI governance. Retrieved from UN News: <https://news.un.org/en/story/2025/09/1165898>

Makhortykh, M., & Mann, H. (2024). AI and the Holocaust: rewriting history? The impact of artificial intelligence on understanding the Holocaust. Retrieved from <https://unesdoc.unesco.org/ark:/48223/pf0000390211>

Mandate of the Special Rapporteur on the negative impact of unilateral coercive measures on the enjoyment of human rights. (2025). Guiding principles on sanctions, business and human rights. Commentary. Retrieved from <https://www.ohchr.org/sites/default/files/documents/issues/ucm/commentary-gpssbhr-2025.pdf>

Mandate of the Special Rapporteur on the negative impact of unilateral coercive measures on the enjoyment of human rights. (2025). Guiding principles on sanctions, business and human rights . Retrieved from <https://www.ohchr.org/sites/default/files/documents/issues/ucm/events/international-conf-sanctions-business-hr/gps-sanctions-business-hr.pdf>

Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America) (Merits) (International Court of Justice June 27, 1986).

Mishra, V. (2025, September 24). United Nations. Retrieved from AI must not decide humanity's fate, UN chief warns Security Council: <https://news.un.org/en/story/2025/09/1165942>

Nations, U. (1984, November 23). Status of the International Convention on the Suppression and Punishment of the Crime of Apartheid, A/Res/39/19. Retrieved from United Nations: <https://docs.un.org/en/A/RES/39/18>

OECD. (2019). OECD AI Principles. Retrieved from <https://www.oecd.org/en/topics/ai-principles.html>

OECD. (2025). Overview and methodology of the AI Incidents and Hazards Monitor. Retrieved from <https://oecd.ai/en/incidents-methodology>

Office for Disarmament Affairs . (2025). Developments in the field of information and telecommunications in the context of international security. Retrieved from United Nations: <https://disarmament.unoda.org/en/our-work/emerging-challenges/developments-field-information-and-telecommunications-context>

Open-ended working group on security of and in the use of information and communications technologies 2021–2025. (2022). Developments in the field of information and telecommunications in the context of international security, A/77/275. Retrieved from <https://docs.un.org/en/A/77/275>

Open-ended working group on security of and in the use of information and communications technologies 2021–2025. (2023). Developments in the field of information and telecommunications in the context of international security, A/78/265. Retrieved from <https://docs.un.org/en/A/78/265>

Open-ended working group on security of and in the use of information and communications technologies 2021–2025. (2024). Developments in the field of information and telecommunications in the context of international security, A/79/214 . Retrieved from file:///C:/Users/user/Downloads/A_79_214-EN.pdf

Pashentsev, E., Bazarkina, D., & Sebekin, S. (2024). Malicious Use of AI and Challenges to Psychological Security of BRICS Countries Malicious Use of AI and Challenges to Psychological Security of BRICS Countries Report. Retrieved from https://www.researchgate.net/publication/379815218_Malicious_Use_of_AI_and_Challenges_to_Psychological_Security_of_BRICS_Countries_Malicious_Use_of_AI_and_Challenges_to_Psychological_Security_of_BRICS_Countries_Report

Protocol additional to the Geneva Conventions of 12 August 1949, and relating to the protection of victims of international armed conflicts (Protocol I). (1977, June 8). Retrieved from United Nations Treaty Collection: <https://treaties.un.org/pages/showdetails.aspx?objid=08000002800f3586>

Puscas, I. (2023, October 12). AI and International Security. Understanding the risks and Paving the Path for Confidence-Building Measures. Retrieved from UNIDIR: https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI-international-security_understanding_risks_paving_the_path_for_confidence_building_measures.pdf

Qian, Y., Siau, K., & Nah, F. (2024, June). Societal impacts of artificial intelligence: Ethical, legal, and governance issues. Societal impacts, 3, 1-5. Retrieved from <https://pdf.sciencedirectassets.com/782283/1-s2.0-S2949697723X00034/1-s2.0-S2949697724000055/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX-2VjEAKaCXVzLWVhc3QtMSJIMEYCIQDibBKV2fCAS5ZpLK21ba47WsZ-bU4U4WOaOmBsDAGR9RgIhAIdEa32pFMehPJdBNAbVMGn32P%2F84c-vcyN3WsDzdCV>

Quelch, I. (2023). Red Cross France tackles AI image fakery with latest campaign. Retrieved from NewDigitalAge: <https://newdigitalage.co/advertising/red-cross-france-not-generated-by-ai-ogilvy/>

Redacción Razones de Cuba. (2024, June 12). Conversando con la IA sobre la situación económica en Cuba (I). Retrieved from <https://razonesdecuba.cu/conversando-con-la-ia-sobre-la-situacion-economica-en-cuba-i/>

Regional Comprehensive Economic Partnership. (2020, November 15). Retrieved from UNCTAD: <https://investmentpolicy.unctad.org/international-investment-agreements/treaties/bilateral-investmenttreaties/4935/rcep-2020->

Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence . (2024, June 13). Retrieved from EUR-lex: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

Ridings, P. (2024). Due Diligence in International Law. Retrieved from International Law Commission: <https://legal.un.org/ilc/reports/2024/english/annex2.pdf>

Safety and security risks of generative artificial intelligence to 2025 (Annex B). (2025). Retrieved from Department for Science, Innovation and Technology: <https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/safety-and-security-risks-of-generative-artificial-intelligence-to-2025-annex-b>

Sami, W. (n.d.). Modern Diplomacy. Retrieved from The Perilous Role of Artificial Intelligence and Social Media in Mass Protests: <https://moderndiplomacy.eu/2024/12/07/the-perilous-role-of-artificial-intelligence-and-social-media-in-mass-protests/>

SCO. (2025, September 1). Tianjin Declaration of the Shanghai Cooperation Organization. Retrieved from DiPLO: <https://www.diplomacy.edu/resource/tianjin-declaration-of-the-shanghai-cooperation-organization/>

Security Council Report. (2023, December 19). Retrieved from Arria-formula Meeting on “Artificial Intelligence: Its Impact on Hate Speech, Disinformation and Misinformation”: <https://www.securitycouncilreport.org/whatsinblue/2023/12/arria-formula-meeting-on-artificial-intelligence-its-impact-on-hate-speech-disinformation-and-misinformation.php>

Security Council Report. (2025, April 4). Retrieved from Arria-formula Meeting Concept Note: Harnessing safe, inclusive, trustworthy AI for the maintenance of international peace and security: <https://www.securitycouncilreport.org/un-documents/document/arria-formula-meeting-concept-note-harnessing-safe-inclusive-trustworthy-ai-for-the-maintenance-of-international-peace-and-security.php>

Technology and society. (2021). Retrieved from Bias and Discrimination in AI: A Cross-Disciplinary Perspective: <https://technologyandsociety.org/bias-and-discrimination-in-ai-a-cross-disciplinary-perspective/>

The Digital Economy Partnership Agreement (DEPA). (2020, June 12). Retrieved from New Zealand Ministry of Foreign Affairs and Trade: <https://isomer-user-content.by.gov.sg/166/9f2ab40f-0ad1-4925-b8d5-03bdb2dda5eb/Digital%20Economy%20Partnership%20Agreement.pdf>

Ministry for Foreign Affairs of Japan <https://www.mofa.go.jp/mofaj/files/100792866.pdf> Threat. (n.d.). Retrieved from Collins Dictionary: https://www.collinsdictionary.com/dictionary/english-thesaurus/threat#threat__1

Threat to peace . (n.d.). Retrieved from The United Nations Terminology Database: <https://unterm.un.org/unterm2/en/view/48737729-e97b-4f07-a86b-ca2d-963c64af>

Toledo, M. C. (2025). Threats, risks and challenges to national securit. Retrieved from Instituto Español de Estudios Estratégicos : https://www.defensa.gob.es/documents/2073105/2383976/amenazas_riesgos_y+desafios_a_la_seguridad_nacional_2025_dieeem03_eng.pdf/10468d78-7b52-6c8f-1d51-27aa1bae1234?t=1759217806364

Townley, S. (n.d.). The Rise of Risk in International Law. Retrieved from Chicago Journal of International Law: <https://cjl.uchicago.edu/print-archive/rise-risk-international-law>

UN AI Advisory body. (2024). Retrieved from Governing AI for humanity : https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

UN General Assembly. (2024, July 1). Enhancing international cooperation on capacity-building of artificial intelligence, A/Res/78/311. Retrieved from <https://docs.un.org/en/A/RES/78/311>

UN Secretary-General's . (2019). Declaration of digital interdependence High-level Panel on Digital Cooperation. Retrieved from United Nations: <https://www.un.org/digital-emerging-technologies/content/about>

UN Security Council. (1993, May 25). Resolution 827 (1993). Retrieved from United Nations: [https://docs.un.org/en/S/RES/827\(1993\)](https://docs.un.org/en/S/RES/827(1993))

UN Security Council. (1994, November 8). Resolution 955 (1994). Retrieved from United Nations: [https://docs.un.org/en/S/RES/955\(1994\)](https://docs.un.org/en/S/RES/955(1994))

UN Security Council. (2024). Pact of the Future. Recommendations for Security Council Action. Retrieved from https://www.securitycouncilreport.org/atf/cf/%7B65BFCF9B-6D27-4E9C-8CD3-CF6E4FF96FF9%7D/chapter_2_report.pdf

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Retrieved from UNESDOC: <https://unesdoc.unesco.org/ark:/48223/pf0000381137>

United Nations. (2004). Retrieved from A more secure world: Our shared responsibility, Report of the High-level Panel on Threats, Challenges and Change: https://www.un.org/peacebuilding/sites/www.un.org.peacebuilding/files/documents/hlp_more_secure_world.pdf

United Nations. (2015). Retrieved from Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, A/70/174: <https://docs.un.org/en/A/70/174>

United Nations. (2023, July 18). Retrieved from Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight, in First Debate on Artificial Intelligence: <https://press.un.org/en/2023/sgsm21880.doc.htm>

United Nations Charter. (1945, June 26). San Francisco.

United Nations Convention against Cybercrime; Strengthening International Cooperation for Combating Certain Crimes Committed by Means of Information and Communications Technology Systems and for the Sharing of Evidence in Electronic Form of Serious Crimes. (2024, December 24). Retrieved from United Nations Treaty Collection: https://treaties.un.org/doc/Treaties/2024/12/20241224%2001-27%20PM/Ch_XVIII_16.pdf

United Nations Convention on the Law of the Sea. (1982, December 10). Retrieved from United Nations Treaty Collection: <https://treaties.un.org/pages/showdetails.aspx?objid=0800000280043ad5>

United Nations General Assembly . (2024, September 22). The Pact for the Future, A/RES/79/1. Retrieved from United Nations: <https://docs.un.org/en/A/RES/79/1>

United Nations General Assembly. (1985, December 16). Economic and social consequences of the armaments race and its extremely harmful effects on world peace and security, A/RES/40/150. Retrieved from United Nations: <https://docs.un.org/en/A/RES/40/150>

United Nations General Assembly. (1987, December 7). Importance of the universal realization of the right of peoples to self-determination and of the speedy granting of independence to colonial countries and peoples for the effective guarantee and observance of human rights. Retrieved from United Nations: <https://digitallibrary.un.org/record/152786?v=pdf#files>

United Nations General Assembly. (1989, December 11). Implementation of the Declaration on the Granting of Independence to Colonial Countries and Peoples. Retrieved from United Nations: <https://docs.un.org/en/A/RES/44/101>

United Nations General Assembly. (1989, December 11). Implementation of the Declaration on the Granting of Independence to Colonial Countries and Peoples, A/RES/44/101. Retrieved from United Nations: <https://docs.un.org/en/A/RES/44/101>

United Nations General Assembly. (2015, December 17). Torture and other cruel, inhuman or degrading treatment or punishment, A/RES/70/146. Retrieved from United Nations: <https://docs.un.org/en/A/RES/70/146>

United Nations General Assembly. (2024). Global digital compact, A/79/L.2. Получено из United Nations: https://www.un.org/global-digital-compact/sites/default/files/2024-09/Global%20Digital%20Compact%20-%20English_0.pdf

United Nations General Assembly. (2024, December 24). Resolution adopted by the General Assembly, A/RES/79/243. Retrieved from <https://docs.un.org/en/a/res/79/243>

United Nations General Assembly. (2024, September). The Pact for the Future. Retrieved from United Nations: <https://docs.un.org/en/a/res/79/1>

United Nations General Assembly. (2025, August 26). Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance. Retrieved from United Nations: <https://docs.un.org/en/A/RES/79/325>

United Nations Global Compact. (2025, April 30). Artificial Intelligence and the Sustainable Development Goals: Operationalizing technology for a sustainable future. Retrieved from United Nations Global Compact: <https://unglobalcompact.org/compactjournal/artificial-intelligence-and-sustainable-development-goals-operationalizing>

United Nations Regional Centre for Western Europe. (2025, September 26). The artificial intelligence we want. Retrieved from <https://unric.org/en/the-artificial-intelligence-we-want/>

United Nations Security Council. (2023, June 14). United Nations. Retrieved from Resolution 2686 (2023): <https://www.securitycouncilreport.org/atf/cf/%7B65BFCF9B-6D27-4E9C-8CD3-CF6E4FF96FF9%7D/s-res-2686.pdf>

United Nations Security Council. (2025). Consideration of the provisions of Chapter V-II of the Charter. Retrieved from Actions with Respect to Threats to the Peace, Breaches of the Peace, and Acts of Aggression: <https://main.un.org/securitycouncil/en/content/repertoire/actions#rel1>

UNSC. (2014-2015). Part VII. Actions with respect to threats to the peace. Retrieved from United Nations: https://www.un.org/securitycouncil/sites/www.un.org/securitycouncil/files/part_vii_2014-2015.pdf#page=6

UNSC. (2016-2017). Part VII. Actions with respect to threats to the peace,. Retrieved from United Nations: https://www.un.org/securitycouncil/sites/www.un.org/securitycouncil/files/part_vii_2016-2017.pdf#page=6

UNSC. (n.d.). Actions with Respect to Threats to the Peace, Breaches of the Peace, and Acts of Aggression. Retrieved from United Nations: <https://www.un.org/securitycouncil/content/repertoire/actions>

Updated concept for a convention of the United Nations on ensuring international information security : Annex to the letter dated 15 May 2023 addressed to the Secretary-General, A/77/894. (2023). Retrieved from United Nations: <https://docs.un.org/en/A/77/894>

Updated Concept of the Convention of the United Nations on Ensuring International Information Security. (2021). Retrieved from [https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-__\(2021\)/ENG_Concept_of_UN_Convention__on_International_Information_Security_Proposal_of_the_Russian__Federation.pdf](https://docs-library.unoda.org/Open-Ended_Working_Group_on_Information_and_Communication_Technologies_-__(2021)/ENG_Concept_of_UN_Convention__on_International_Information_Security_Proposal_of_the_Russian__Federation.pdf)

Verush, A. (2018). Relation of categories “risk”, “challenge”, “threat” in the theory of national security (rus). Scientific Works of the Republican Institute of the Higher School. Philosophical-Humanitarian Sciences, 17, 24-30.

Vienna Convention on Consular Relations. (1963, April 26). Retrieved from United Nations Treaty Collection: https://treaties.un.org/Pages/ViewDetails.aspx?src=IND&mtdsg_no=III-6&chapter=3

Vienna Convention on Diplomatic Relations. (1961, April 18). Retrieved from United Nations Treaty Collection: https://treaties.un.org/pages/viewdetails.aspx?src=treaty&mtdsg_no=iii-3&chapter=3&clang=_en

Vienna Convention on the Law of Treaties. (1969, May 23). United Nation, Treaty Series, 1155, 331. Vienna.

Völk, K., & Nguyen, K. (2025). BBC news. Retrieved from BBC reveals web of spammers profiting from AI Holocaust images: <https://www.bbc.com/news/articles/ckg4xjk1g1xo>

Wet, E. D., & Wood, M. (2022, January). Peace, Threat to the. Retrieved from Oxford Public International Law: <https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e374>

Wood, M., & Sender, O. (2024). Identification of Customary International Law. Oxford: Oxford University Press.

World Economic Forum. (2025, September). Advancing Responsible AI Innovation: A Playbook. Insight report. Retrieved from https://reports.weforum.org/docs/WEF_Advancing_Responsible_AI_Innovation_A_Playbook_2025.pdf

World Health Organization . (2023). Regulatory considerations on artificial intelligence for health. Retrieved from <https://iris.who.int/server/api/core/bitstreams/ad62580f-540f-4e36-b957-e7f2946ae1fb/content>

Yampolskiy, R. (2024). *AI Unexplainable, Unpredictable, Uncontrollable*. New York: Routledge.

Yousif, N. (2025, July 8). Imposter used AI to pose as Marco Rubio and contact foreign ministers. Retrieved from BBC: <https://www.bbc.com/news/articles/crrqkyjjewno>

