DOI: https://doi.org/10.24833/0869-0049-2024-4-85-105

Исследовательская статья
УДК: 341
Поступила в редакцию: 19.07.2024
Принята к публикации: 26.11.2024

Екатерина Александровна МАРТЫНОВА
Национальный исследовательский университет «Высшая школа экономики»
Мясницкая ул., д. 20, Москва, 101000, Российская Федерация
eamartynova@hse.ru
ORCID: 0000-0002-8995-4462

«ПУТЕШЕСТВИЕ В ПОСЛЕЗАВТРА»: РЕГУЛИРОВАНИЕ КИБЕРОПЕРАЦИЙ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В КОНТЕКСТЕ ПРИМЕНЕНИЯ СИЛЫ

ВВЕДЕНИЕ. *Искусственный интеллект (далее – ИИ) может значительно укрепить системы информационной безопасности государств, а также послужить дополнительным техническим средством совершения злонамеренных действий в так называемом киберпространстве. Стремясь получить конкурентное преимущество в цифровой сфере, государства начали инвестировать в оборонительные и наступательные автономные кибервозможности для защиты своих интересов и сдерживания потенциальных противников, что подстегнуло рост числа межгосударственных кибeропераций. Статья посвящена не только проблемам применения существующих норм международного права в ситуациях злонамеренного использования ИИ государствами, но и процессу интерпретации этих норм различными субъектами и через эту интерпретацию кристаллизации общего (или, по крайней мере, сближающегося) понимания их применимости. Таким образом, в данной работе рассматривается путь к пониманию того, как нормы международного права о применении силы действуют в отношении кибeропераций с использованием ИИ.*

МАТЕРИАЛЫ И МЕТОДЫ. *Настоящее исследование основано на работах как российских, так и зарубежных специалистов в области права международной информационной безопасности, а также на анализе документов и материалов групп правительственных экспертов под эгидой Организации Объединенных Наций и позиций государств. Помимо общенаучных методов (анализ, синтез, индукция и дедукция), в исследовании применяется теория транснационального правового процесса, которая рассматривает текущие дискуссии по соответствующим вопросам на различных площадках и в целом кооперацию различных акторов в процессе формирования правил ответственного использования ИИ государствами посредством взаимодействия, толкования и интернализации интерпретированных правовых идей и практик во внутренние правовые системы.*

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ. *Международные усилия по разработке универсального свода правил ответственного поведения государств в киберпространстве пока не увенчались успехом. Проанализированная история обсуждения допустимых действий государств* *в киберпространстве позволяет предположить, что дискуссия, посвященная ИИ, в обозримом будущем будет развиваться вне контекста разработки всеобъемлющего международного договора. Вместо этого правовой ландшафт применения ИИ, судя по всему, будет формироваться на основе инструментов «мягкого права» и инициатив частного сектора, что повлечет за собой фрагментацию толкования и практики государств.*

ОБСУЖДЕНИЕ И ВЫВОДЫ. *Усложнение межгосударственных кибeропераций технологиями ИИ поднимает дополнительные вопросы о действии международного права, в частности его норм о применении силы, в отношении киберинцидентов с использованием ИИ. Поскольку государства стремятся разработать и приобрести смертоносные автономные системы вооружений для поддержания стратегического паритета, что может дестабилизировать глобальную безопасность и повысить риск эскалации конфликтов, развертывание подобных систем и участие государств в кибeрoперациях с использованием ИИ способно привести к очередной гонке вооружений — на этот раз с применением ИИ. Это и другие политические и этические соображения говорят о целесообразности ограничения свободы действий государств в использовании ИИ. Однако на сегодняшний день стимулы для стран НАТО, Китая и России договориться о международном юридически обязательном документе, пресекающем использование ИИ в злонамеренных целях, представляются иллюзорными. Исходя из истории дискуссий о применении международного права в киберпространстве и разработке правил ответственного поведения государств, использующих информационно-коммуникационные технологии, можно предположить, что соответствующие дискуссии об ИИ, скорее всего, будут проходить вне рамок разработки международного договора. Таким образом, дальнейший анализ развития этого вопроса потребует изучения того, как транснациональные нормы, например те, что возникают из инструментов «мягкого права», практики государств и инициатив частного сектора, будут формировать международно-правовой ландшафт применения ИИ.*

КЛЮЧЕВЫЕ СЛОВА: *искусственный интеллект, присвоение поведения, кибeроперации, киберпространство, применение силы, ответственность государства, транснациональный правовой процесс*

ДЛЯ ЦИТИРОВАНИЯ: Мартынова Е.А. 2024. «Путешествие в послезавтра»: регулирование кибeропераций с использованием искусственного интеллекта в контексте применения силы. - *Московский журнал международного права.* № 4. С. 85–105. DOI: https://doi.org/10.24833/0869-0049-2024-4-85-105

Данная работа подготовлена в рамках стипендии для молодых ученых в области международного права, предоставленной Центром международных и сравнительно-правовых исследований, Москва, Россия.

Название статьи «Путешествие в послезавтра» отсылает к одноименному научно-фантастическому роману Роберта Шекли (1962).

Автор заявляет об отсутствии конфликта интересов.

CYBERSPACE AND INTERNATIONAL LAW

DOI: https://doi.org/10.24833/0869-0049-2024-4-85-105

Research article
UDC: 341
Received 19 July 2024
Approved 26 November 2024

Ekaterina A. MARTYNOVA
National Research University Higher School of Economics
20, Myasnitskaya St., Moscow, Russian Federation, 101000
eamartynova@hse.ru.
ORCID: 0000-0002-8995-4462

"JOURNEY BEYOND TOMORROW": NAVIGATING REGULATION OF AI-POWERED CYBER OPERATIONS IN THE REALM OF THE USE OF FORCE

INTRODUCTION. *Artificial intelligence (AI) can significantly strengthen cybersecurity systems of States, as well as serve an additional technical means for malicious actions in the so-called cyberspace. Recognizing this, States have started investing in defensive and offensive autonomous cyber capabilities to protect their interests and deter potential adversaries; this has further fuelled the increase in inter-State cyber operations as nations seek to gain a competitive edge in the digital realm. This paper focuses not only on the problems of applying existing norms of international law to situations of malicious use of AI by States, but also on the process of these norms' interpretation by different actors and, through this interpretation, crystallization of a common (or at least converging) understanding of their applicability. More specifically, this paper examines the path to understanding of how the norms on the use of force apply to AI-enabled cyber operations.*

MATERIALS AND METHODS. *The present study is based on the works of both Russian and foreign specialists on the law of international information security, as well as analysis of documents and materials of groups of governmental experts under the auspices of the United Nations and the positions of States. In addition to general scientific methods (analysis, synthesis, induction and deduction), the theory of transnational legal process is applied to this study, which considers ongoing discussions of relevant issues on various platforms, and, more generally, the interaction of various actors regarding the formation of a pool of rules for responsible use of AI by States through interaction, interpretation and internalization of the interpreted legal ideas and practices into the domestic legal systems.*

RESEARCH RESULTS. *The international efforts to develop a universal set of rules for responsible State behaviour in cyberspace have enjoyed rather modest success. The analysed history of cyber-related debate suggests that the AI-focused discussion for the foreseeable future will progress outside the area of developing a comprehensive treaty framework. Instead, the legal landscape of AI applications will appear to emerge from soft law instruments and private sector initiatives, which would lead to fragmentation of interpretation and State practice.*

DISCUSSION AND CONCLUSIONS. *The complication of inter-State cyber operations by AI technology raises additional questions about the application of international law, in particular its norms on the use of force, to AI-powered cyber incidents. The deployment of lethal autonomous weapons systems and commitment of AI-powered cyber operations could potentially lead to another – this time, AI – arms race, as nations seek to develop and acquire these systems to maintain strategic parity. This could destabilize global security and increase the risk of conflict escalation. This and other political and ethical considerations argue in favor of limiting the discretion of States in the use of AI. However, to date, the incentives for NATO States, China and Russia to agree on an international binding instrument limiting the use of AI for malicious purposes appear illusory. One could argue that corresponding discussions on AI will probably take place outside of the development of an international treaty, given the historical debate surrounding the application of international law in cyberspace and the development of norms governing responsible States behaviour in the use of information and communication technologies. Further analysis of this development, thus, will require examining how transnational norms, such as those emerging from soft law instruments, customary practices, and private sector initiatives, will shape the international legal landscape of the AI application.*

KEYWORDS: *artificial intelligence, attribution, cyber operation, cyberspace, use of force, State responsibility, transnational legal process*

FOR CITATION: Martynova E.A. "Journey Beyond Tomorrow": Navigating Regulation of AI-Powered Cyber Operations in the Realm of the Use of Force. – *Moscow Journal of International Law.* 2024. No. 4. P. 85–105. DOI: https://doi.org/10.24833/0869-0049-2024-4-85-105

This paper was prepared within the framework of the Scholarship for Young International Law Researchers provided by the International and Comparative Law Research Center, Moscow, Russia.

The "Journey Beyond Tomorrow" in the title of this paper refers to the science fiction novel (1962) of the same name by Robert Sheckley.

The author declares the absence of conflict of interest.

1. Introduction

Over the last two decades, there has been a notable upsurge in the number of inter-State cyber operations taking place worldwide[1]. Most of them are classified as acts of espionage[2] as traditional methods of intelligence gathering have been enhanced by cyber capabilities, but some incidents (and their number is increasing as technology advances) have consequences in the physical space, including causing damage to individuals[3], companies[4], infrastructure facilities[5] and computer networks[6]. The relative anonymity and deniability offered by cyberspace make it an attractive operational domain for States to gain strategic advantages without the fear of immediate retaliation. Amid the growing reliance on digital techniques and interconnectedness of cyber infrastructure systems States' vulnerability to cyberattacks is increasing. This vulnerability is not limited to military or government networks but extends to sectors such as energy, finance, transportation, and healthcare.

The introduction of artificial intelligence (AI) technology[7] and the increased autonomy (and therefore, unpredictability and unreliability [Stroppa

[1] According to the Council of Foreign Relations that has tracked significant cyber operations since 2005 starting with one cyber operation perpetrated by China that year, following with at least 30 State-sponsored cyber operations in 2015, 76 cyber operations sponsored by States in 2019, and 125 incidents in 2022. URL: www.cfr.org/cyber-operations/ (accessed date: 15.07.2024).

[2] One of the recent cases is phishing campaign reportedly conducted in March 2023 by the Pakistani cyberespionage group APT 36 against the Defence Research and Development Organisation – Indian governmental agency involved in research and development of sensitive defence technologies used by the Indian Armed Forces. See: Notorious SideCopy APT Group Sets Sights on India's DRDO – *Cyble*. 21 March 2023. URL: https://cyble.com/blog/notorious-sidecopy-apt-group-sets-sights-on-indias-drdo (accessed date: 15.07.2024).

[3] E.g., attacks of targeted individuals – human rights activists and university professors by APT37 (aka RedEyes, ScarCruft, and Reaper), a hacking group allegedly sponsored by North Korea. See: RedEyes Group Wiretapping Individuals (APT37) – *ASEC*. 21 June 2023. URL: https://asec.ahnlab.com/en/54349/ (accessed date: 15.07.2024).

[4] E.g., attacks against subsidiary companies of multinational corporations based in the US and East Asia, including industrial, technology, media, electronics, and telecommunications companies, with further access to the parent companies' network by BlackTech – cyber actors linked to the People's Republic of China. See: Press release "CISA, NSA, FBI and Japan Release Advisory Warning of BlackTech, PRC-Linked Cyber Activity" – *CISA*, 27 September 2023. URL: www.cisa.gov/news-events/news/cisa-nsa-fbi-and-japan-release-advisory-warning-blacktech-prc-linked-cyber-activity (accessed date: 15.07.2024).

[5] E.g., targeting from late 2021 to mid-2022 of US critical infrastructure including seaports, energy companies, transit systems, and a major US utility and gas entity by Mint Sandstorm – a group associated with the Iranian government. See: Microsoft Threat Intelligence. "Nation-State Threat Actor Mint Sandstorm Refines Tradecraft to Attack High-Value Targets". 18 April 2023. URL: www.microsoft.com/en-us/security/blog/2023/04/18/nation-state-threat-actor-mint-sandstorm-refines-tradecraft-to-attack-high-value-targets/ (accessed date: 15.07.2024).

[6] E.g., an attack by a North Korean hacking group of the networks of Seoul National University Hospital in 2021 as a result of which hackers gained access to the personal medical records of hundreds of thousands patients. See: Reddy S., "North Korean Hackers Stole 830K People's Data in Attack on Seoul Hospital: ROK" – *NK News.* 10 May 2023. URL: www.nknews.org/2023/05/north-korean-hackers-stole-830k-peoples-data-in-attack-on-seoul-hospital-rok/ (accessed date: 15.07.2024).

[7] Despite the ubiquitous nature of AI discussions lately, there is no consistent 'official' definition of AI. In some cases, the technical descriptions offered by computer scientists are not suitable for legal analysis, for example when AI is defined in terms of an 'algorithm', which in turn requires a separate definition and understanding of the social meaning and legal content. For the review of different approaches to define AI for the purposes of legal studies, refer, e.g. to [Lee 2022:6-8].
This paper does not aim to go deeper into the search for a correct definition of AI; the now traditional approach to define AI in the narrow and broad senses suits the purposes of this study: "More specifically, that there is a key difference between narrow AI such as translation services, chatbots, and autonomous vehicles and general AI – "self-learning systems that can

2023:3]) of inter-State cyber operations further enhances this vulnerability. One can assume with a certain degree of confidence that the use of AI by States will expand. This is evidenced by numerous forecasts by cybersecurity experts[8], and direct declarations of States[9], coupled with an unprecedented increase in investments in research and development of AI-powered capabilities[10].

The advancement of AI and machine learning (ML) technologies[11] has given rise to a discourse surrounding the potential risks posed by autonomous (including, AI-driven) cyber operations to international peace and security, as well as the right to self-defence against such actions. The corpus of research that has emerged in recent years addresses how current international legal norms are applied to novel circumstances involving employment of AI technologies by States, including concerns such as violation by autonomous cyber capabilities of the principles of sovereignty and non-intervention [Schmitt 2020:554‑558]; interpretation of the precautionary principle of international law with respect to "lethal AI" [Garcia 2018:335, 338]; unilateral and collective autonomous measures of self-help against malicious cyber-enabled activities[12]. The Russian doctrine in this area focuses mainly on the challenges for international humanitarian law (hereinafter, the **IHL**) related to robotisation and autonomisation of weapon systems [e.g., Morkhat 2017; Proskurina, Khokhlova, Safin 2020] and the application of AI in the context of armed conflict [Chernyavsky, Sibileva 2020].

The focus of this paper, however, is not only on the problems of applying existing norms of international law to situations of malicious use of AI by States, but also on the process of these norms' interpretation by different actors and, through this interpretation, crystallization of a common (or at least converging) understanding of their applicability. More specifically, this paper examines the path to understanding of how the norms on the use of force apply to AI-enabled cyber operations and why States might agree to obey them in this context. The theory of transnational legal process is applied to this study, which considers ongoing discussions of relevant issues on various platforms, and, more generally, the interaction of various actors regarding the formation of a pool of rules for responsible use of AI by States.

The scope of research needs to be specified according to the aims defined. First, the application of AI and specifically cyber operations based on the AI technology will not be analyzed from the standpoint of international human rights law and IHL. Thus, this study focuses on *jus ad bellum* body of international law. Second, with respect to the question of AI employment in cyber operations, the scope *ratione personae* is limited to the activities that are conducted or

learn from experience with humanlike breadth and surpass human performance on all tasks". General AI raises broader existential concerns, such as how to align the goals of such a system with our own to prevent catastrophic outcomes, but general AI remains a technology still to be developed in the distant future.' (Meltzer J.P. The Impact of Artificial Intelligence on International Trade – *Brookings*. 13 December 2018. URL: www.brookings.edu/research/the-impact-of-artificial-intelligence-on-international-trade/#footnote-1 (accessed date: 15.07.2024)).

[8] UNIDIR: The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations. Report No. 7. 2017. URL: https://unidir.org/files/publication/pdfs/autonomous-weapon-systems-and-cyber-operations-en-690.pdf (accessed date: 15.07.2024).

[9] Thus, the Cyber Security Strategy for Germany of 2021 indicates the aim to achieve 'the highest possible level of IT security for AI systems' and examine continually "the opportunities for using AI systems to protect (government) IT systems". See: Cyber Security Strategy for Germany 2021, at 46-47. URL: www.bmi.bund.de/EN/topics/it-internet-policy/cyber-security-strategy/cyber-security-strategy-node.html (accessed date: 15.07.2024). In a similar mode, the US National Cybersecurity Strategy of 2023 sets forth as a strategic objective focus on technologies "that will prove decisive for U.S. leadership in the coming decade" including AI, and states that such effort "will facilitate the proactive identification of potential vulnerabilities, as well as the research to mitigate them". See: The White House: Fact Sheet: Biden-Harris Administration Announces National Cybersecurity Strategy at 25 (2 March 2023). URL: www.whitehouse.gov/briefing-room/statements-releases/2023/03/02/fact-sheet-biden-harris-administration-announces-national-cybersecurity-strategy/ (accessed date: 15.07.2024).

[10] The projected growth of the AI market is up to 1.8 trillion US dollars by 2030: "Artificial Intelligence (AI) Market Size Worldwide in 2021 with a Forecast until 2030 (in million U.S. dollars)". – *Statista*. URL: www.statista.com/statistics/1365145/artificial-intelligence-market-size/ (accessed date: 15.07.2024).

[11] Machine learning is either classified as a subfield of AI or a distinct field and refers to the creation of digital systems that become more proficient at a particular task over time through experience. See: Brundage M. et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Oxford: Future of Humanity Institute. 2018. URL: https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf (accessed date: 15.07.2024).

[12] Stroppa M. 2023. Autonomous Cyber Capabilities and Unilateral Measures of Self-Help against Malicious Cyber Operations. – *NATO CCDCOE Publication*. URL: https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-and-unilateral-measures-of-self-help-against-malicious-cyber-operations/ (accessed date: 15.07.2024).

sponsored by States. This entails analyzing the problem of attribution of malicious activities using AI to the allegedly sponsoring State.

This paper is structured as follows. Section Two provides a brief overview of the development of the international debate on autonomous weapon systems and cyber operations. Section Three describes key characteristics of AI-powered malicious cyber operations relevant for further legal analysis and discusses application of the UN Charter norms on the use of force and self-defence to such cyber incidents; the matter of attribution of such activities to the organizing State is also discussed. In Section Four, the focus shifts from an analysis of the possible application of the lex lata international law to looking at the process of forming relevant norms on inter-State AI-enabled cyber operations; an attempt is made to get a view of this 'journey beyond tomorrow' through the prism of transnational legal process theory. The final Section summarizes the discussion.

2. Evolution of the international debate on autonomous weapons and cyber operations

Concerns about autonomous weapons systems, also known as lethal autonomous weapons systems (hereinafter, the **LAWS**) and lethal autonomous robotic systems (hereinafter, the **LARS**), or "killer robots", have been addressed since 2014 under the auspices of the Convention on Certain Conventional Weapons (hereinafter, the **CCW**). With States, international organizations, and civil society participating,[13] the CCW formed a Group of Governmental Experts (hereinafter, the **GGE on LAWS**) in 2016 with the express purpose of debating and resolving issues pertaining to LAWS. Enhancing knowledge of the technical, legal, ethical, and operational facets of LAWS is the goal of the discussions. Its focus is on the potential impact of these systems on armed conflicts, including questions of human control, accountability, compliance with IHL, and the potential consequences for civilians. In 2019 the GGE on LAWS adopted Guiding Principles[14] reaffirming the application of IHL to all weapon systems, including autonomous ones[15], underlining human responsibility for decisions on the use of LAWS "since accountability cannot be transferred to machines"[16], emphasizing consistently and in relation to several aspects of the development, testing and use of LAWS the need for analysis of compliance with international law, in particular IHL[17], and pointing out that LAWS should not be anthropomorphized in domestic policies[18].

Although a legally binding agreement on LAWS has yet to be reached, the GGE on LAWS has facilitated significant discourse among States and non-State actors, and has contributed to raising awareness at the international level about the necessity for regulation in this domain. It can be argued that settling of the principles that are quite general in nature is "costless" for States in the sense that no firm obligations are assumed in contrast to a universal legally binding instrument, reaching agreement on which in the nearest future seems to be an almost impossible task. That said, the approach of the GGE on LAWS is to foster transparency of discussion, promote dialogue, and build consensus among States on how best to address the challenges posed by autonomous weapons systems. As will be shown further in this paper, the repeated interaction between actors, during which the interpretation of legal norms and concepts occurs, is an essential part of the transnational legal process and can have norm-forming potential even in the absence of the conclusion of an international treaty as a result of such interaction.

[13] Thus, in the latest session of the GGE on LAWS in May 2023, apart from representatives of the States – contracting parties to the CCW, participated representatives of the ICRC, UNIDIR, non-governmental organizations including Amnesty International, Future of Life Institute and Human Rights Watch, as well as members of the academic community representing Universität Bremen, University of Cambridge, University of Oxford, University of Queensland and some other institutions. See: Report of the 2023 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. 24 May 2023. CCW/GGE.1/2023/2. URL: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_2_Advance_version.pdf (accessed date: 15.07.2024).

[14] Annex III to the Final report of the Meeting of the High Contracting Parties to CCW. 13 December 2019. CCW/MSP/2019/9. URL: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/343/64/PDF/G1934364.pdf?OpenElement (accessed date: 15.07.2024).

[15] Ibid. P. 10. Para (a).

[16] Ibid. P. 10. Para (b).

[17] Ibid. P. 10. Paras (c), (d), (e), and (h).

[18] Ibid. P. 10. Para (i).

The development and deployment of LAWS has also given rise to significant debates and concerns within the international scientific and humanitarian community. These have included calls for a complete ban on LAWS for ethical and legal reasons [Asaro 2012:694-703]. Thus, expanding the use of killer robots could lead to dehumanization of war and erosion of moral accountability in warfare by allowing machines to make life and death decisions without human involvement [Sassòli 2014]. These voices, however, are becoming weaker as the use of LAWS and LARS nestle down as an integral part of any armed conflict involving a State with a sufficient level of technological development and cyber capabilities. At the same time, proponents of a balanced approach to the use of LAWS by States raise concerns about their compliance with IHL, in particular the principles of distinction, proportionality and military necessity: LAWS may not possess the ability to differentiate between combatants and civilians, potentially leading to unlawful attacks [Anderson, Reisner, Waxman 2014:401]. The question of legal responsibility also is complicated with autonomy of weapons if humans are not directly involved in decision-making processes and, thus, requires preventive security framework based on the precautionary principle of international law [Garcia 2018: 338; Chernyavsky, Sibileva 2020:235].

That said, States' interest in autonomy is not limited to LAWS and LARS – autonomisation of technology is becoming ubiquitous and is also penetrating the field of cyber operations. Technological advancements in strengthening autonomy in digital and physical systems are happening rapidly. As was already mentioned, the international discussions on autonomous weapons systems and autonomous cyber capabilities are taking place independently of each other and asynchronously. It is in some ways ironic that the issue of regulating autonomy-enhancing technologies is discussed primarily in relation to conventional weapons in the context of the CCW, while the operation of such technologies and their consequences manifest themselves on a significant scale in the digital space, and not (only) in the real world, and in peacetime, not (only) in time of armed conflict. Consequently, the issue of autonomous (in particular, AI-powered) cyber operations, which is of paramount importance in evaluating the influence of AI on international security, is frequently either overlooked or dismissed. While there is significant focus on systems like LARS and LAWS, these represent only a fraction of the various potentially hostile applications that can be enabled by AI.

The international discussion under the UN aegis about the risks of inter-State cyber operations for the international security began long before the discussion of LAWS in the context of the CCW. The use of information and communication technologies (hereinafter, the **ICTs**) in the context of international security is addressed since 1998, when Russia introduced a draft resolution on "information security" in the First Committee of the UN General Assembly. The potential threats to global security and stability posed by the development of cyber capabilities have been examined by six Groups of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (hereinafter, the **UN GGE**) since 2004–2005 . "International law and in particular the United Nations Charter, is applicable and is essential to maintaining peace and stability and promoting an open, secure, peaceful and accessible ICT environment", the UN GGE confirmed in its 2013 milestone report[20]. Although there were ups and downs in the further work of the UN GGE (in particular, the group was unable to adopt a consensus report in 2015 due to differences in the positions of States on the issues of the application of IHL and the right to self-defence to cyber incidents), the sixth and final GGE, which met between 2019 and 2021, reached a consensus on report[21] that further contributed to the understanding of responsible behaviour of States in cyberspace. In December 2018, the General Assembly established an Open-Ended Working Group (hereinafter, the **OEWG**) to address the existing and emerging threats in the ICTs environment and the application of international law in cyberspace[22]. The OEWG, open to all UN Member States, held meet-

[19] See the factsheet: Developments in the Field of Information and Telecommunications in the Context of International Security. URL: https://disarmament.unoda.org/ict-security/ (accessed date: 16.07.2024).

[20] UNGA: Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security. 24 June 2013. UN Doc A/68/98 (GGE Report 2013).

[21] UNGA: Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security. 14 July 2021. UN Doc A/76/135 (GGE Report 2021).

[22] UNGA: Resolution "Developments in the field of information and telecommunications in the context of international security". 11 December 2018. UN Doc A/RES/73/27.

ings with various stakeholders including industry, civil society and academia and adopted a consensus report in March 2021, which was endorsed by the General Assembly[23]. In 2020, a new five-year OEWG was established to continue addressing ICTs security[24], and in 2022, a resolution entitled "Programme of action to advance responsible State behaviour in the use of information and communications technologies in the context of international security" was adopted by the General Assembly requesting a report from the Secretary-General on advancing responsible State behaviour in the use of ICTs[25]. The use of AI in the malicious actions of States in cyberspace is outside the mandate of the UN GGE and OEWG, but has received some, albeit limited, attention in the scientific literature in connection with the discussion of so-called autonomous cyber capabilities[26].

Interestingly, in the GGEs on LAWS and the UN GGEs there was practically no overlap of technical and legal specialists among the participants. The discussions have been held in different formats, most conversations within the GGEs on LAWS and the UN GGEs have been centred around a particular technology as though the technologies in question are developing independently. The conceptual apparatus used by the groups also differs. Accordingly, a few terminological clarifications are in order. Unlike relatively established terminology of LARS and LAWS in the studies of the warfare robotization, definitions in the field of studying inter-State cyber operations are just being coined. In the studies under the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE) auspices, the term 'autonomous cyber capabilities' is used with clarification that autonomy means, in the first place, autonomy from a human operator[27], which can technically be provided by both AI and other programming means such as pre-defined tasks or sequences of actions. Such capabilities can function without AI, but incorporating AI can further enable them to make independent decisions or adjust behavior in changing circumstances; moreover, AI can be used in an assistance role in software operated by humans. In this context, the element of AI application is taken out of the analysis: "AI is an enabler for autonomy but neither synonymous with it nor a prerequisite for it"[28].

François Delerue in his seminal *Cyber Operations and International Law* employs the term "autonomous cyber operations" describing them as those which "once activated, can select and engage targets without further intervention by a human operator" [Delerue 2020:158]. He comes to the conclusion that the autonomy of a cyber operation does not affect its attribution to a State, since ultimately every cyber operation is created, programmed and launched by human operators [Delerue 2020:162].

Empirical studies of cases where AI has been used to carry out cyber operations also use different terminology and are based on different methods for assessing the role of AI in the cyber incident. The literature review conducted by the Swedish Defence Research Agency revealed 19 cases of reported "artificially intelligent cyberattacks" as of March 2020[29]. In accordance with the division of the cyberattack anatomy into stages, introduced by the authors of the study, AI was used and there is technical readiness for its use at all five stages: reconnaissance, access and penetration, internal reconnaissance and lateral movement, command and control, as well as exfiltration and sanitation[30], but primarily on the early phases[31]. A more recent study shows that the majority of "AI-driven cyberattack'" techniques were identified in the access and penetration phase, followed by the use in exploitation and command and

[23] UNGA: Open-ended working group on developments in the field of information and telecommunications in the context of international security. Final Substantive Report. 10 March 2021. A/AC.290/2021/CRP.2 (OEWG Report 2021).

[24] UNGA: Resolution "Developments in the field of information and telecommunications in the context of international security". 4 January 2021. UN Doc A/RES/75/240.

[25] UNGA: Resolution "Programme of action to advance responsible State behaviour in the use of information and communications technologies in the context of international security". 12 December 2022. UN Doc A/RES/77/37.

[26] Liivoja R., Naagel M. and Väljataga A. 2019. Autonomous Cyber Capabilities under International Law. – *NATO CCDCOE Publications*. URL: https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law/ (accessed date: 10.07.2024).

[27] Ibid. P. 9.

[28] Ibid. P. 10.

[29] Zouave E. et al. 2020. Artificially Intelligent Cyberattacks. – FOI-R--4947—SE. URL: www.statsvet.uu.se/digitalAssets/769/c_769530-l_3-k_rapport-foi-vt20.pdf at 8 (accessed date: 10.07.2024).

[30] Ibid. P. 17.

[31] Ibid. P. 37.

control phases[32]. It is worth emphasizing that there is very little empirical data in relation to inter-State AI-enabled incidents due to the predominantly classified status of such information.

To date, fully autonomous AI-enabled cyber operations have not been recorded. However, many of those phenomena that were previously the figment of imagination of science fiction writers are now reality. The rapid development of technology and the associated threats to international security require proactive analysis, particularly taking into account that cyber operations are increasingly driven by AI[33]. Accordingly, the next section of this paper describes the key characteristics of the AI-powered inter-State cyber operations and provides a general overview of the application of the norms on use of force to them.

3. AI-powered cyber operations from the use of force perspective

For the purposes of further discussion, the term 'AI-powered cyber operation' (hereinafter, the **AIpCO**) is used, referring to the cases when a State or a non-State actor whose conduct is attributed to a State employs AI capabilities to execute a cyberattack or defensive cyber-enabled actions without direct human intervention at least at some stages of the relevant activity. The AIpCOs range from simple, predefined actions to complex decision-making processes based on AI algorithms. An AIpCO might have the ability to spread itself through autonomous decision-making, adapting its tactics based on the specific characteristics of the infected system, similar to the WannaCry worm attack but with a more intelligent and adaptable approach[34]. AI-powered malware is potentially able to learn from contextual information, imitate trustworthy system components, exploit vulnerabilities, and evade detection, thus enabling it to cause extensive harm[35]. AIpCOs can expand the existing threats of spear phishing attacks, which is obtaining data or initiating action from the attack's target by deceiving them with a seemingly reliable front[36]. Moreover, they empower novel types of malicious cyber activities, involving for instance deepfakes or voice mimicking. Among the most extreme scenarios, an AIpCO in conjunction with social engineering techniques, such as fabricating a military order, is carried out or AI in another form is abused to enable nuclear weapons if nuclear control systems were digitalized and inadequately secured[37].

Among the well-known and thoroughly studied inter-State cyber incidents, the Stuxnet attack can be classified (with some degree of caution) as an early AIpCO. Discovered in June 2010, Stuxnet was a highly sophisticated cyberattack that specifically targeted industrial control systems and programmable logic controllers (PLCs) used in Iran's nuclear facilities. Stuxnet utilized advanced techniques, including the use of multiple zero-day vulnerabilities and rootkit capabilities, to infect its targets and remain undetected[38]. It also employed complex algorithms and AI techniques to make decisions based on the specific characteristics of the targeted systems. Once inside the system, Stuxnet would manipulate the PLCs, causing them to malfunction and operate outside of their intended parameters. This resulted in physical damage to centrifuges used for uranium enrichment, leading to a significant slowdown in Iran's nuclear program. The level of complexity suggests that the Stuxnet worm was likely developed and deployed by a nation-State with significant resources. It has been widely speculated that the United States and Israel collaborated on the development of Stuxnet[39], although the attack was not attributed legally to any of

[32] Guembe B. et al. 2022. The Emerging Threat of Ai-driven Cyber Attacks: A Review. – *Applied Artificial Intelligence.* Vol. 36. Issue 1. e2037254. URL: https://doi.org/10.1080/08839514.2022.2037254 (accessed date: 10.07.2024).

[33] Miller M. NATO Prepares for Cyber War. – *Politico*. 3 December 2022. URL: www.politico.com/news/2022/12/03/nato-future-cyber-war-00072060 (accessed date: 10.07.2024).

[34] The Next Paradigm Shift AI-Driven Cyber-Attacks. – *DarkTrace Research White Paper.* 2021. URL: www.oixio.ee/sites/default/files/the_next_paradigm_shift_-_ai_driven_cyber_attacks.pdf, at 2 (accessed date: 10.07.2024).

[35] Ibid. P. 3.

[36] Brundage et al (n 11). P. 18.

[37] For the description of cyber-nuclear conflicts scenarios refer, e.g., to Stoutland P. O. and Pitts-Kiefer S. Nuclear Weapons in the New Cyber Age. – *Nuclear Threat Initiative.* 26 September 2018. URL: www.nti.org/analysis/reports/nuclear-weapons-cyber-age (accessed date: 10.07.2024).

[38] Kaspersky. Stuxnet Explained: What It Is, Who Created It and How It Works. URL: www.kaspersky.com/resource-center/definitions/what-is-stuxnet (accessed date: 10.07.2024).

[39] Kushner D. The Real Story of Stuxnet. – *IEEE Spectrum.* 19 January 2024. URL: https://spectrum.ieee.org/the-real-story-of-stuxnet (accessed date: 15.07.2024).; Council on Foreign Relations. Stuxnet. URL: www.cfr.org/cyber-operations/stuxnet (accessed date: 15.07.2024).

these States. Stuxnet was unique in its ability to sabotage physical infrastructure through cyber means. It demonstrated the potential for cyberattacks to cause real-world damage and highlighted the growing importance of cybersecurity in critical infrastructure protection. Stuxnet was widely regarded groundbreaking also as an example of a cyberattack able to conduct autonomously and carry out complex actions without human intervention[40]. Stuxnet, thus, "appears to be the first autonomous weapon with an algorithm, not a human hand, pulling the trigger"[41].

Before proceeding to the analysis of the norms on use of fore potentially applicable to AIpCOs such as Stuxnet and more advanced ones, it is helpful to address their properties relevant for the analysis of the legal consequences of the AIpCOs commission, particularly those distinguishing AIpCOs from other types of cyber operations.

a. Certain characteristics of AIpCOs relevant for further discussion[42]

The first characteristic that lies on the surface is that AI and ML technologies are employed to automate various aspects of an AIpCO. For the AIpCOs involved in the cybersecurity systems this includes tasks such as threat detection, incident response (including, automated "hack-back"), vulnerability management, and security analytics. Leveraging AI and ML algorithms enables analyzing large volumes of data in real-time, identifying patterns, and detecting anomalies that may indicate potential cyber threats or attacks. Additionally, these technologies can continuously learn and adapt to new threats and attack techniques, improving their effectiveness over time. The AIpCOs committed for malicious purposes leverage AI and ML to compromise security of the target and the integrity of its computer networks. In the cases when a State is a target of an AIpCO, its digital security (e.g. provision of public services which are digitalized), physical security (if an AIpCO reaches the level of an armed attack) and political security (e.g., in the cases of interference in political processes including elections, or targeted disinformation campaigns) are put at risk[43].

This description also highlights the second characteristic of AIpCOs – their dual-use nature, that is applicability both for cybersecurity and hostile actions [Taddeo, McCutcheon, Floridi 2019:557]. Defensive AIpCOs focus on preventing, detecting, and responding to cyber threats and attacks employing AI. This includes activities such as implementing security controls, monitoring network traffic for anomalies, conducting vulnerability assessments, and incident response – "hack back"[44]. Offensive operations, on the other hand, involve actively targeting and compromising adversary networks, systems, or data. Such operations can include activities such as reconnaissance, exploitation, and disruption of adversary's capabilities[45]. These operations are typically carried out in a controlled manner and may involve intelligence gathering, penetration testing, or even offensive cyber warfare in certain contexts. Both offensive and defensive AIpCOs are essential components of a comprehensive State cybersecurity strategy.

Finally, it is necessary to note such a feature of AIpCOs as autonomy which refers to the degree to which AI systems can independently make decisions and carry out actions within the cyber operation without human intervention. This level of autonomy can vary, ranging from systems that require human approval for every action ("human-in-the-loop"), then the systems operating autonomously but under human supervision ("human-on-the-loop"), to those that can operate with minimal or no human oversight ("human-out-of-the loop")[46]. Advances in ML

[40] Kaspersky (n 38).

[41] Healey J. Stuxnet and the Dawn of Algorithmic Warfare. – *Huffington Post*. 16 April 2013. URL: www.huffingtonpost.com/jason-healey/stuxnet-cyberwarfare_b_3091274.html (accessed date: 11.07.2024).

[42] Since the author is not a specialist in the field of computer science, the summary of characteristics given in the text does not claim to be a comprehensive and technically accurate description. The purpose of this section is to point out those features of AIpCOs that are important for further legal analysis.

[43] Brundage et al (n 11). P. 10.

[44] Tammet T. 2021. Autonomous Cyber Defence Capabilities. – *Autonomous Cyber Capabilities under International Law*. Liivoja R., Väljataga A. (eds). (NATO CCDCOE Publications). At 39. 42-45. URL: https://ccdcoe.org/uploads/2021/05/Autonomous-Cyber-Capabilities-under-International-Law.pdf (accessed date: 11.07.2024).

[45] The Use of Artificial Intelligence in Cyber Attacks and Cyber Defense. 2 September 2023. URL: https://secureops.com/blog/ai-offense-defense/ (accessed date: 15.07.2024).

[46] Garcia E.V. 2019. Artificial Intelligence, Peace and Security: Challenges for International Humanitarian Law. – *Cadernos de Política Exterior nº 8*. Instituto de Pesquisa de Relações Internacionais (IPRI). Brasilia. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3595340 (accessed date: 11.07.2024).

and AI algorithms have enabled cyber operations to become more autonomous allowing them to adapt and evolve their strategies based on real-time feedback and changing circumstances. The autonomy of AIpCOs offers several advantages, which might be especially relevant for defensive operations. In particular, it allows for faster response times, as AI can detect and mitigate threats in real time, without the need for human intervention, as well as adapts to the changing circumstances and corrects actions due to behaviour of adversaries without additional input from the human operator[47]. It can also handle a larger volume of attacks, improving the overall security posture. Moreover, AI can identify complex attack techniques and vulnerabilities that may be difficult for humans to detect. However, there are challenges associated with the autonomy of AIpCOs, and the lack of human oversight raises concerns about liability and ethics. In situations where AI systems make decisions that have significant consequences, such as launching offensive operations or making critical infrastructure decisions, the potential for unintended consequences or malicious use due to lack of the situational awareness and political understanding of reality becomes a concern [Stroppa 2023:2].

b. Application of international law on the use of force to AIpCOs

The growing sophistication and autonomy of AIpCOs raise significant concerns about their potential impact, moral implications and legality, including in relation to the corpus of international law on the use of force. This calls for a review of the extent to which AIpCOs can be conducted in accordance with Article 2(4) of the UN Charter, which prohibits the use of force, and Article 51, which concerns the right of self-defence.

The two-threshold approach outlined in the UN Charter serves as the foundation for the implementation of international law principles pertaining to the use of force. This approach encompasses the duality of the "use of force" and "an armed attack" (Article 2(4) of the UN Charter). Furthermore, the "scale and effects" doctrine, as articulate by the International Court of Justice (hereinafter, the **ICJ**) in the *Nicaragua* case[48], provides additional support for this approach. One of the most frequently cited sources on the application of the norms of international law in the context of cyberspace – the *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* [Schmitt 2017] (hereinafter, the **'Tallinn Manual 2.0'**) – suggests that a cyber operation constitutes a use of force, and thus its commission violates Article 2(4), if its "scale and effect" are comparable to a kinetic attack that rises to the level of the use of force[49]. It is evident that the UN Charter lacks an appropriate definition of what would constitute force and (or) an armed attack, particularly, in cyberspace. Furthermore, there is a lack of clarity regarding the threshold of cyberattacks that would invoke Article 51.

It is sometimes claimed in the literature that the notion "force", in line with the ICJ approach in the Dispute regarding *Navigational and Related Rights (Costa Rica v. Nicaragua)*[50] should have an evolving meaning, as other generic terms employed in international treaties of unlimited duration [Roscini 2014:46-47]. Thus, certain States have identified different approaches to the evolutionary interpretation of the term "force", including the so-called "effects-based", "instrument-based" and "target-based" approaches.

Those States that adopt the "effects-based" approach go so far as to apply the evolutionary interpretation of the term "force" in Article 2(4), thereby including cyber operations without physical effects within its scope. France suggested adopting a non-exhaustive list of criteria for the assessment of whether a 'cyber operation without physical effects' can be characterized as the use of force, which criteria include "the circumstances prevailing at the time of the operation, such as the origin of the operation and the nature of the instigator (military or not), the extent of intrusion, the actual or intended effects of

[47] McFarland T. 2021. The Concept of Autonomy. (n 44). P. 23.

[48] ICJ: Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America). Merits Judgment of 27 June 1986. – *I.C.J. Rep.* 1986. Para 195.

[49] Tallinn Manual 2.0. Rule 69.

[50] ICJ: Dispute regarding Navigational and Related Rights (Costa Rica v. Nicaragua). Judgment, 2009. – *ICJ Rep.* 213. Para 66: "[W]here the parties have used generic terms in a treaty, the parties necessarily having been aware that the meaning of the terms was likely to evolve over time, and where the treaty has been entered into for a very long period or is 'of continuing duration', the parties must be presumed, as a general rule, to have intended those terms to have an evolving meaning".

the operation or the nature of the intended target"[51]. According to a declaration made by the Dutch Minister of Foreign Affairs, a cyber operation that has a substantial financial or economic impact could be considered a use of force[52]. The UK Cyber Primer, while acknowledging the necessity for a cyber operation to cause "the same or similar effects as a kinetic attack", makes a clarification in a footnote that permits such a qualification for attacks, like "a sustained attack against the UK banking system, which could cause severe financial damage to the state leading to a worsening economic security situation for the population"[53]. This may be indicative of a desire on the part of certain States (it should be noted that their number is limited) to extend the scope of the internationally prohibited "use of force" through the use of domestic efforts that could be considered an indication of State practice and *opinio juris*. Other States, while espousing the "effect-based" approach, are circumspect in applying it to cyber operations devoid of physical damage. For instance, Italy stated that "[w]hile it is generally accepted that cyber operations resulting in material damage can constitute a use of force", it considers "the qualification of cyber operations which merely cause loss of functionality a controversial one"[54].

Another analytical approach involved for the assessment of the use of force prohibition to cyber-enabled activities is the "instrument-based" approach [Roscini 2014:47]. It is predicated on the general and broad acknowledgement in national perspectives that any use of force, regardless of the type of weapon, might violate the prohibition of the use of force[55]. It is underpinned by the permissibility of the consequential use of the analogy with kinetic attacks (causing deaths, injury or the destruction of physical objects). The applicability of this approach is questioned by some States, in particular by Germany, which "shares the view that with regard to the definition of "use of force", emphasis needs to be put on the effects rather than on the means used"[56].

Finally, a "target-based" approach can be involved to substantiate that a cyberattack constitutes use of force if it is conducted against national critical infrastructure of a victim State [Roscini 2014:47]. This approach is endorsed by Estonia in combination with the "effects-based" approach: "A cyber operation that targets critical infrastructure and results in serious damage, injury or death, or a threat of such an operation, would be an example of use of force"[57]. Israel extends it by providing practical examples: "Hacking into the computers of the railroad network of another State and programming the controls in a manner that is expected to cause a collision between trains can amount to use of force"[58], referring also to the effects of the attack alongside its target. Overall, the "target-based" approach, if employed in isolation, carries the risk of expansive application, as it labels any cyber operation targeting a critical national infrastructure as a use of force, even if it causes only nuisance or merely gathers information. Besides, there is no consensus on what exactly constitutes "critical national infrastructure", which creates further ambiguity. The Russian national regulation, thus, broadens the definition of "critical information infrastructure objects" by listing the sectors in which

[51] Ministère des Armées. France: International Law Applied to Operations in Cyberspace. P. 7. URL: https://www.defense.gouv.fr/sites/default/files/ministere-armees/Droit%20international%20appliqu%C3%A9%20aux%20op%C3%A9rations%20dans%20le%20cyberespace.pdf (accessed date: 15.11.2024).

[52] Letter from the Minister of Foreign Affairs to the President of the House of Representatives on the International Legal Order in Cyberspace. 5 July 2019. Appendix: International Law in Cyberspace. P. 4.

[53] UK Ministry of Defence: Cyber Primer. 2nd ed. 2016. Annex 1A – *International Law aspects*. P. 12.

[54] Italian Ministry for Foreign Affairs and International Cooperation: Italian position paper on "International law and cyberspace. 2021. P. 8. URL: https://www.esteri.it/mae/resource/doc/2021/11/italian_position_paper_on_international_law_and_cyberspace.pdf (accessed date: 15.11.2024).

[55] ICJ: Legality of the Threat or Use of Nuclear Weapons. Advisory Opinion. 1966. – *ICJ Rep.* 226. Para 39.

[56] Germany. On the Application of International Law in Cyberspace: Position Paper. P. 6. URL: https://www.auswaertiges-amt.de/blob/2446304/32e7b2498e10b74fb17204c54665bdf0/on-the-application-of-international-law-in-cyberspace-data.pdf (accessed date: 15.11.2024).

[57] Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by states submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security. 13 July 2021. UN Doc. A/76/136. P. 26.

[58] Schondorf R. Israel's Perspective on Key Legal and Practical Issues Concerning the Application of International Law to Cyber Operations. – *EJIL:TALK!* 9 December 2020. P. 399. URL: https://www.ejiltalk.org/israels-perspective-on-key-legal-and-practical-issues-concerning-the-application-of-international-law-to-cyber-operations (accessed date: 15.11.2024).

relevant information systems and networks operate. These sectors include healthcare, science, transport, communications, energy, banking and other financial market sectors, the fuel and energy complex, nuclear energy, defence, rocket and space industry, mining, metallurgy and chemical industry[59]. Thus, physical objects of information infrastructure and telecommunication networks used to organize interaction between them constitute the concept of "critical national infrastructure", attack on which can potentially constitute use of force if Russia applied the "target-based" approach.

As shown by the various methods examined, the severity of the attack, its immediacy, directness, invasiveness, measurability of effects, military nature, State involvement and the alleged legality of the attack are just some of the quantitative and qualitative factors that can be used to determine whether a cyber operation crosses the line into the use of force. However, the assessment of certain of these qualitative factors, including intensity of the operation, is hardly possible by means of autonomous AI algorithms, which requires pre-programming or operator control to comply with restrictions on the use of force [Stroppa 2023:3]. Whether AI will in principle be able to assess the legality of an attack is a question that remains open. Furthermore, the potential consequences of autonomous cyber operations as a means of force extend beyond immediate physical damage. The interconnected nature of cyberspace means that attacks can have cascading effects, impacting critical infrastructure, economic stability, and even human lives. Evaluating the potential collateral damage and long-term consequences of such operations becomes essential in assessing their legality.

The notion of the 'armed attack' defined by the ICJ as "the most grave forms of the use of force"[60] requires a nuanced approach when applied for qualification of actions in cyberspace. The international group of experts – authors of the Tallinn Manual 2.0 suggested that a cyber operation reaches the level of an armed attack if it kills or seriously injures several persons or significantly damages or destroys property[61]. In the context of cyber operations, this could theoretically mean situations where critical infrastructure, such as power grids or transport systems, is targeted, resulting in physical damage or significant disruption, or causing death, for example by an attack on a hospital network. However, the matter of qualification of hostile activities in cyberspace as an armed attack remains highly controversial, which is confirmed, inter alia, by the lack of uniformity in the States positions within the UN GGE working sessions and the fact that, as a consequence, this issue remained unresolved in the last report of the UN GGE[62]. The uncertainty of States as to whether a cyber operation can constitute an armed attack is manifested, for example, in the persistent mantra of analysing each incident "on a case-by-case basis"[63].

The issue of applicability of the right to self-defence against cyber operations became a bone of contention in the work of the UN GGEs. Both the 2013 and 2015 reports acknowledged the role of the UN Charter in promoting international peace, but the right to self-defence (Article 51) sparked intense arguments in every UN GGE session leading up to their adoption and reached its peak within the work of the UN GGE on the fifth report in 2016–2017. Western nations pushed to explicitly include self-defence as a response to cyberattacks, but some other countries resisted. Russia, China and Cuba were particularly opposed to this idea. Some expressed the concern that extending Article 51 of the UN Charter to cyberattacks could lead to a constant cycle of conflict in the digital realm. Particularly, the position of Cuba was based on the assertion that the draft text of

[59] Федеральный закон «О безопасности критической информационной инфраструктуры Российской Федерации» [Federal Law "On the Security of Critical Information Infrastructure of the Russian Federation"]. No. 187-FZ dated 26 July 2017. Art. 2(8). (In Russ.)

[60] *Nicaragua v US.* Para. 191.

[61] Tallinn Manual 2.0. Rule 71. Para 8.

[62] See Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security. UN Doc A/76/135 (14 July 2021). Para. 71(f).

[63] See, e.g.: African Union Peace and Security Council: Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace. 29 January 2024. P. 6: "Whether a particular cyber operation constitutes a use of force or amounts to an armed attack should be determined on a case-by-case basis"; Compendium. P. 69-70. Position of Norway: "A cyber operation may constitute use of force or even an armed attack if its scale and effects are comparable to those of the use of force or an armed attack by conventional means. This must be determined based on a case-by-case assessment having regard to the specific circumstances".

the UN GGE report "aimed to establish equivalence between malicious use of ICTs and the concept of 'armed attack', as provided for in Article 51"[64].

The content of measures constituting self-defence also raises the question of their nature. To date, consensus has been formed in the national positions of the States that have expressed themselves on this issue that self-defence in response to a cyberattack can be expressed by both cyber and kinetic means; conversely, cyber means can be used in self-defence against a kinetic attack[65]. The logic behind this attitude was expressed by Poland which indicated that "[d]eprivation of the right to respond to such a cyberattack with kinetic means could render the self-defence right illusory when the perpetrator of an armed attack is little dependent on its functioning in cyberspace"[66].

In order for AI to exercise the right of self-defence under Article 51 of the UN Charter, it must be able, first, to classify hostile actions, whether kinetic or cyber, as an "armed attack" and, second, to respond within the predetermined parameters of self-defence, namely necessity and proportionality[67]. Furthermore, a response in the form of hacking back necessitates an irrefutable attribution of the initial attack to the perpetrator to circumvent the potential for the deployment of force against a third State and the concomitant liability for the responding State. The use of various masking and mimicry technologies complicates technical attribution, not to mention the complexity of legal attribution of activities to a State, which is primarily a political decision, and therefore requires a broader contextual approach than that which can be provided by AI.

c. Attribution of AIpCOs to States

The question of the possibility and procedure for attributing actions committed during an AIpCO to a State is connected, on the one hand, with the limits of the autonomy of such an operation (which were discussed earlier), and on the other – with the broader and highly controversial issue of potential legal personality of AI.

Addressing the latter aspect, it is worth noting that the Articles on Responsibility of States for Internationally Wrongful Acts (hereinafter, the **ARSIWA**)[68] operates in terms "person", "a group of persons" and "entity" listing the actors whose conduct can be attributed to a State[69]. According to the ILC commentaries to ARSIWA[70], the term "person or entity", although used in a broad sense, includes a natural or legal person[71] but not a technology. The term "group of persons" is intended to emphasize the fact that the attributable conduct "may be that of a group lacking separate legal personality but acting on a *de facto* basis"[72]. In this paradigm centered on personality, it seems that actions committed within an AIpCO with a high level of autonomy cannot in principle be attributed to a State at least until AI technology is given legal personality. Two possible approaches can be preliminary discussed.

The first approach lies in the plane of recognizing the AI legal personality, whereupon the concept of "person" for the purposes of applying the law of State responsibility will include not only natural and legal persons, but also AI as a technology or algorithm, depending on how it is defined for the purposes of

[64] Declaration by Miguel Rodríguez, Representative of Cuba, at the Final Session of Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security. New York. 23 June 2017.

[65] See, e.g.: International law and cyberspace. Finland's national positions. P. 7. URL: https://um.fi/documents/35732/0/KyberkannatPDF_EN.pdf/12bbbbde-623b-9f86-b254-07d5af3c6d85?t=1603097522727 (accessed date: 15.11.2024).

[66] Ministry of Foreign Affairs of Poland, The Republic of Poland's Position on the Application of International Law in Cyberspace at 5. 29 December 2022. URL: https://www.gov.pl/web/diplomacy/the-republic-of-polands-position-on-the-application-of-international-law-in-cyberspace (accessed date: 15.11.2024).

[67] *Nicaragua v US*. Paras. 176, 194. On proportionality in the *jus as bellum* for autonomous cyber capabilities see Margulies P. 2021. Autonomous Cyber Capabilities, Proportionality, and Precautions. (n 44). P. 162-165.

[68] UNGA: Res. 56/83. Articles on Responsibility of States for Internationally Wrongful Acts 12 December 2001. UN Doc. A/RES/56/83.

[69] Article 4(2) ARSIWA: "An organ includes any person or entity which has that status in accordance with the internal law of the State". The term "person or entity" is also used in Article 5 and 7. The term "group of persons" is used in Articles 8 and 9.

[70] International Law Commission: Articles on the Responsibility of States for Internationally Wrongful Acts, with Commentaries. 2001. UN Doc. A/56/10. (The ARSIWA w. Commentaries).

[71] ARSIWA w. Commentaries, commentary (12) to Article 4.

[72] Ibid. Commentary (9) to Article 8.

designating its personality. In this case, AI would be considered a person whose actions might, in theory, be attributed to a State under the general grounds for attributing conduct to a State, specifically as directed or controlled by a State under Article 8 ARSIWA. The discussion about the (im)practicability and (im)possibility of recognizing the legal personality of AI is by now quite robust and relates to the issues of liability of AI in general, for instance, in the context of self-driving cars, as well as in other areas, including intellectual property rights [Čerka, Grigienė, Sirbikytė 2017; Chesterman 2020]. In the context of the law of State responsibility, it can be advocated that at least limited personhood shall be accorded to AI once it reaches the level of autonomous decision making [Stahl 2006:211]. Curiously, in August 2022 the US Department of the Treasury's Office of Foreign Assets Control (OFAC) imposed sanctions on the so-called virtual currency mixer "Tornado Cash"[73] which technically is a computer code, a set of smart contracts on the Ethereum blockchain, and does not fall under the definition of a "person" according to Executive Order No. 13694, on the basis of which the sanctions were applied. However, apart from this exception case, to date there is no evidence of State practice or *opinio juris* to support a supposition that AI or another technology can be treated as a legal subject from the international law perspective[74]. This first approach, at the same time, introduces a wide range of ethical issues related to the anthropomorphization of technology, which cannot be resolved solely by the methods of legal and political science.

The second approach is perhaps more instrumental and consists in recognizing AI exclusively as a tool for performing AIpCOs, regardless of the degree of the algorithm autonomy in making decisions. In this case, the question of conduct attribution refers not to the "State – AI" relation, but to the relation 'State – natural or legal person' that programmed, launched or supervised an AIpCO, even if the degree of autonomy of such operation reached the level "human-out-of-the-loop". This approach brings the debate about responsibility for inter-State AIpCOs closer to the discussion of responsibility for using LAWS. At the same time, with perhaps a simpler legal implementation compared to the first approach (as it does not create novel legal obstacles), it is associated with certain practical difficulties, some of which are already observed while attempting to attribute "old-fashioned" cyber operations without AI, and which get even more complicated with the addition of an AI element to them. These difficulties relate both to the application of specific grounds of attribution (particularly, the State control over the conduct of non-State actors under Article 8 ARSIWA) and a duty to reach any of the standards of proof applicable in international law [Roscini 2015:248]. Thus, starting with the early well-known cases of cyber operations targeted against States, such as a cyberattack against Estonia in 2007[75], States have been resorting quite willingly to accusations and condemnation. However, the transition from "naming and shaming" [Finnemore, Hollis 2020] to official attribution is complicated by legal and political reasons: namely, the need to prove the ground of attribution, that is the link between the attack perpetrator and the State, and to disclose evidence. Instead, States tend to use rather vague language about their belief in another State sponsoring malicious cyber activity: so, the UK rated the probability of the Russian military intelligence service (GRU) perpetrating a cyberattack against Georgia in 2019 as "almost certain" (95 % + probability)[76], though no evidence proving attribution was provided. Until recently, a situation in which a State would admit involvement in malicious actions in cyberspace against another State, much less proactively declare them, was unthinkable. However, since October 2023, the Main Intelligence Directorate of the Ukrainian Ministry of Defense has openly accepted responsibility for at least five episodes of cyberattacks, the most plausible of which was the statement about an attack on the Russian company IPL Consulting which develops and implements technological solutions for managing processes in heavy

[73] U.S. Department of the Treasury: Press release "Treasury Sanctions Notorious Virtual Currency Mixer Tornado Cash". 8 August 2022. URL: https://home.treasury.gov/news/press-releases/jy0916 (accessed date: 11.11.2024).

[74] Liivoja, Naagel and Väljataga (n 26). P. 32.

[75] Estonian Links Moscow to Internet Attack. – *The New York Times.* 18 May 2007. URL: www.nytimes.com/2007/05/18/world/europe/18estonia.html (accessed date: 15.11.2024).

[76] Foreign & Commonwealth Office: Press Release "UK Condemns Russia's GRU over Georgia Cyber-Attacks". 20 February 2020. URL: www.gov.uk/government/news/uk-condemns-russias-gru-over-georgia-cyber-attacks (accessed date: 15.11.2024).

industry[77]. Probably, official statements on behalf of a Ukrainian government agency are designed for some informational effect, and exclude the possibility of any real legal consequences in terms of responsibility, which once again emphasizes the highly politicized nature of the issue of attribution.

In case of AIpCOs, complications are added to the matter of attribution from both sides: substantive and evidentiary[78]. The introduction of an additional element of AI into the process of establishing the connection between the individual or individuals responsible for the cyber operation and the State may present a significant challenge in determining the grounds for attribution, namely on the basis of State control over the conduct under Article 8 ARSIWA, if the perpetrator himself does not control the progress (in a worst case scenario, even the launch) of the operation due to the fact that decision-making is transferred to AI. The regime of objective responsibility can probably be applied to any malicious use by a State of AI technically attributable to that State, regardless of its degree of autonomy. However, for the practical application of such a regime, States should agree, not at the level of abstract principles, but in the form of concrete commitments to ensure transparency and traceability of the use of AI. The fundamental achievability of such agreements raises serious doubts, based, among other things, on the entire history of discussions between States about norms on responsible behavior in cyberspace, which shows a low level of trust among participants of this process in relation to each other. In the absence of such commitment on traceability, attribution of the malicious use of AI will require provision of evidence, which in its turn potentially might lead to the disclosure of sensitive information about the victim-State's cyber defence capabilities, including AI systems employed for the task of national cybersecurity. This will lead to the same situation that already exists with regard to conventional cyber operations, where States are extremely reluctant to provide evidence of their accusations against alleged attackers for fear of compromising the secrecy of data about their own cyber capabilities[79].

4. "Journey beyond Tomorrow": formation of norms on responsible use of AI by States as transnational legal process

This section of the study is based on the methodology of Transnational Legal Process (hereinafter, the **TLP**), also known as the "New" New Haven School of International Law. TLP is a conceptual framework for understanding international law and a school of thought that provides a comprehensive approach to the study of how international legal norms are created by private and public actors, applied, interpreted, enforced and internalized in a globalized world [Koh 2001:313]. The applicability of this methodological approach to researching the process of formatting international (or transnational) legal regulation of AI application in general, and in the context of cyber operations in particular, can be justified by specific features of TLP as a mode of international legal scholarship. At the heart of this approach lies the idea that legal norms are internalized through repeated cycles

[77] Fornusek M. Military Intelligence Claims Cyberattack on IT Company Providing Services to Russian Defense Industry. – *The Kyiv Independent*. 27 January 2024. URL: https://kyivindependent.com/military-intelligence-claims-powerful-cyberattack-on-russian-it-company (accessed date: 11.11.2024).

[78] For a detailed discussion of applying possible grounds of attribution under ARSIWA refer, e.g., to Haataja S. 2021. Attribution in the Law of State Responsibility. – Liivoja and Väljataga (n 44). P. 260.

[79] The question of attribution became a subject matter of national contributions which certain States submitted within the work of UN GGE in 2019–2021. Most of the contributing States expressed quite a cautious position about the standard of attribution of malicious cyber-enabled activities to another State. E.g., Australia noted that "States are entitled, in their sole discretion, and based on their own judgement, to attribute unlawful cyber activities to another State. States should act reasonably when drawing conclusions based on the facts before them". The necessity to reveal evidence is not mentioned. Similar position was expressed by Estonia: "Attribution remains a national political decision based on technical and legal considerations regarding a certain cyber incident or operation. Attribution will be conducted on a case-by-case basis, and various sources as well as the wider political, security and economic context can be considered", as well as Germany which considers that "that there is no general obligation under international law as it currently stands to publicize a decision on attribution and to provide or to submit for public scrutiny detailed evidence on which an attribution is based" and some other States. Contribution of the Russian Federation, on the contrary, called for the avoidance of unsubstantiated accusations and for disclosure of necessary technical evidence. See the "Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution 73/266". 13 July 2021. UN Doc A/76/136*.

of "interaction-interpretation-internalization", where "particular readings of applicable global norms are eventually domesticated into states' domestic legal systems" [Koh 2007:566] and through this compliance of States with international law is promoted. The process is shaped by the interaction of the so-called "agents of internalization", which include non-State actors (so-called transnational norm entrepreneurs), official representatives of States (governmental norm sponsors), transnational issue networks, and interpretive communities [Koh 2007:567-568]. The agents of internalization prompt interaction through which the interpretation or articulation of the global norm relevant to the circumstance takes place. The aim of the initiator might be not (or not only) in the coercion of the counterparty to obey the norm but in internalization of the international norm in the counterparty's domestic system [Koh 2022:113].

The four distinctive features of the transnational legal process, as they are described by Harold Hongju Koh, one the founding fathers of TLP, fully manifest in the process of international AI- and cybernorms formation. First, the TLP as a methodological approach is non-traditional in the sense that it dismantles the traditional dichotomies between international and domestic, public and private [Koh 2001:314]. It is evident that the process of formalising norms on the behaviour of actors in cyberspace and the nascent discussion on AIpCOs does not occur within the confines of either the national or international level. This is because it is neither a matter of regulation that is confined solely to the borders of each particular State nor is it solely a matter of regulation that is confined to operating across their borders. Consequently, the norms and guidelines defining acceptable State conduct in cyberspace with the aim of mitigating the risks posed by cyber activity both within and between States and fostering stability, security, and trust in the digital sphere, can be found in national legislation and at the international level. Domestic norms, included particularly in cybersecurity strategies and doctrines with the EU Cybersecurity Strategy of 2020[80] and the US National Cybersecurity Strategy of 2023[81] among recent instances, establish regulations that outline the rights, responsibilities, and liabilities of individuals, organizations, and institutions concerning cybersecurity. The EU Artificial Intelligence Act represents one of first examples of comprehensive domestic regulation of AI, which introduces certain limitations in application of AI aiming to protect citizens' rights[82]. National laws in this field also typically cover areas such as data protection, privacy, intellectual property, and cybercrime. International norms can be found both in international treaties, including Budapest Convention on Cybercrime and the Draft UN Convention against cybercrime adopted by consensus on 7 August 2024, as well as in an emerging form in sources of soft law such as the UN GGE principles of responsible State behaviour in cyberspace, recommendations of the UN High-level Advisory Body on Artificial Intelligence encapsulated in the Final Report "Governing AI for Humanity"[83] and the "Recommendation on the Ethics of Artificial Intelligence", adopted in 2021 by UNESCO. The formalization of AI and cybersecurity norms also dissolves the private-public divide because they are intended to regulate State-to-State relations (mainly defining the boundaries of legitimate inter-State influence in cyberspace), which makes them a component of "public" international law, and at establishing a procedure for cross-border interaction (for example, with respect to personal data transfer) between non-State actors – that is, in the context of "private" international law.

Second, transnational legal process is non-statist meaning that it involves not only States but also, and sometimes primarily, non-State actors [Koh 2001:314]. The subject of the application of international law in cyberspace, and in particular the debate surrounding the requirement, or perhaps the redundancy, of international legal regulation of the use of AI, is a truly multi-stakeholder one. Furthermore, international cybersecurity norms encompass confidence-building measures (CBMs), which are

[80] Joint Communication to the European Parliament and the Council on the EU's Cybersecurity Strategy for the Digital Decade. 16 December 2020. URL: https://digital-strategy.ec.europa.eu/en/library/eus-cybersecurity-strategy-digital-decade-0 (accessed date: 13.07.2024).

[81] The White House (n 9).

[82] European Parliament. Press release "Artificial Intelligence Act: Deal on Comprehensive Rules for Trustworthy AI". 9 December 2023. URL: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai (accessed date: 11.07.2024).

[83] Governing AI for Humanity: Final Report. 2024. URL: https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf (accessed date: 16.11.2024).

voluntary actions taken by States to increase transparency and reduce the risk of misperception or miscalculation in cyberspace[84]. CBMs include sharing information on national cybersecurity strategies, participating in joint exercises, and establishing hotlines for communication during cyber incidents. In the paradigm of TLP, repeated situations of interaction between State representatives, international organizations and bodies (such as the Internet Governance Forum), expert groups, academia and civil society have the potential to crystallize the norms of international law applicable to the inter-State cyber incidents, including AIpCOs, through the interpretation of existing provisions and facilitation of discussions, negotiations, and consensus-building among nations.

Third, transnational legal process is not static but dynamic [Koh 2001:314]: the norms change and evolve in the cycles of interactions, they move bottom-up from private initiatives to the legislator, and *vice versa,* and also migrate from the national to the inter-State level, and in the opposite direction. Thus, the emerging field of international soft law regarding AI relationships results in the interaction of various sources, such as private agreements (often, market-driven [Chinen 2023:72-106]) made by corporations, regulations set by individual States, the body of international law itself, recommendations of international organizations, and civil society, which predefines formulation of the normative framework through a range of approaches, from voluntary agreements to formal regulations.

Finally, transnational legal process is normative [Koh 2001:314]. Subsequent stages of the interactive cycle of interaction and interpretation internalize these norms and promulgate compliance with them as part of the States' internal normative system and value set. With respect to AIpCOs the regulation decisions are multilevel, and the interaction between the "agents of internalization" leads to informal lawmaking, in particular through interpretations on the international level with further incorporation into national policies[85]. As an example, the AI Policy Observatory was established by the OECD with the purpose of overseeing AI policies and producing documents that have an impact on national policies: the "Recommendation of the Council on Artificial Intelligence" of 2019 (amended in 2023) contains, *inter alia,* recommendations for the States adhering to this document on the development of national policies for AI, including experimental regimes for testing AI systems[86].

Regional and State positions on cyber-related issues, including rule of international law, are anything but uniform, and today this split is more visible than ever. Thus, NATO's 2022 Strategic Concept adopted at the Madrid Summit in June 2022 emphasizes the need for enhancement of the "strategic partnership" between the Alliance and the EU, including in the task of "countering cyber and hybrid threats"[87]. Cooperation contemplates exchange of cyber threat intelligence, joint training and research as well as regulatory convergence. As another vivid and recent example, the new US National Cybersecurity Strategy[88] premises on the opposition of the States designated as authoritarian, which include China, Russia, Iran and North Korea[89], and countries that share the declared commitment of the United States to democratic freedom and human rights. This opposition is repeatedly emphasized throughout the text of the US National Cybersecurity Strategy: it goes as far as naming China, Russia, Iran and North Korea as the main adversaries of the US in cyberspace which "with revisionist intent are aggressively using advanced cyber capabilities to pursue objectives that run counter to our interests and broadly accepted international norms"[90] (although, the US Strategy does not clarify which international norms precisely). On the other side, Russia and China as part of the BRICS, signed the XIV BRICS Summit Beijing Declaration which underscored "the importance of establishing legal frameworks of cooperation among

[84] Healey J. et al. Confidence-Building Measures in Cyberspace. A Multistakeholder Approach for Stability and Security. – *Atlantic Council.* 2014. URL: www.files.ethz.ch/isn/185487/Confidence-Building_Measures_in_Cyberspace.pdf (accessed date: 12.07.2024).

[85] On the review of national AI policy regimes and their typology depending on the State's governance mode see: Filgueiras F. 2022. Artificial Intelligence Policy Regimes: Comparing Politics and Policy to National Strategies for Artifiial Intelligence. – *Global Perspectives.* Vol. 3. Issue 1. URL: https://doi.org/10.1525/gp.2022.32362 (accessed date: 11.07.2024).

[86] OECD: Recommendation of the Council on Artificial Intelligence. Paris: OECD. 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed date: 13.07.2024).

[87] NATO 2022 Strategic Concept, para. 43. URL: www.nato.int/strategic-concept/ (accessed date: 14.07.2024).

[88] The White House (n 9).

[89] Ibid. at 3.

[90] Ibidem.

BRICS countries on ensuring security in the use of ICTs" and acknowledged "the need to advance practical intra-BRICS cooperation through implementation of the BRICS Roadmap of Practical Cooperation on ensuring security in the use of ICTs"[91].

In the context of such open confrontation, are there prospects for a rapprochement, in particular on the issue of the responsible use of AI? Will the US, China, and Russia be prepared to collaborate and restrict the use of AI for malevolent purposes, given their divergent stances but shared aspirations to lead the development of AI?

The history of the UN GGE work shows how States are able to come from an absolute dead end on certain matters if not to consensus, then to compromise, through repeated episodes of interaction and communication. One illustrative example is the discussion of the application of *jus ad bellum,* IHL and the law of State responsibility in the context of cyber incidents. The fifth UN GGE was unable to adopt final reports in 2016–2017 due to concerns expressed by Russia[92] and China[93] and Cuba[94] that the applicability of *jus ad bellum* and IHL may result in the establishment of an "equivalence between the malicious use of [information and communication technologies] and the concept of 'armed attack'"[95] under Article 51 of the UN Charter. This, in turn, could lead to the militarisation of the use and the response to ICTs. In its report of 2021, the UN GGE finally acknowledged the applicability of IHL, specifying though that these norms apply "only in situations of armed conflict"[96], thus leaving the door open for further discussion on which particular cyber operations can be qualified as an "armed conflict". Still, progress in the work from the fifth to sixth UN GGE shows the possibility in principle to achieve certain compromise.

In discussion on the limitation of malicious AI use, including LAWS and employment of autonomous cyber capabilities for hostile actions, analogy is sometimes drawn between regulation of AI and restrictions over specific types of weapons, including nuclear arms control[97]. However, the prospect of agreeing on a hard law international treaty limiting AI similar to weapons on the platform of the UN GGE or the GGE in LAWS is assessed as low, and some argue that other processes could produce more effective formal regulation[98]. Besides, the scepticism that limitation of the malicious use of AI will follow the same path as non-proliferation of nuclear weapons stems from the comprehension that AI today poses a lesser existential threat than nuclear weapons did against the backdrop of the Cuban missile crisis when the talks initiated and eventually led to negotiation of the Treaty on the Non-Proliferation of Nuclear Weapons[99]. And since the stakes are perceived not so high, the motivation to reach agreements with competitors is elusive, which leads some commentators to suggest that achieving a universal international treaty to limit the malicious use of AI is impossible until the world is on the verge of a major AI-caused crisis[100].

[91] XIV BRICS Summit Beijing Declaration. Para. 31. URL: www.gov.br/mre/en/contact-us/press-area/press-releases/xiv-brics-summit-beijing-declaration (accessed date: 11.07.2024).

[92] Response of the Special Representative of the President of the Russian Federation for International Cooperation on Information Security Andrey Krutskikh to TASS' Question Concerning the State of International Dialogue in This Sphere. 29 June 2017. URL: www.mid.ru/en/foreign_policy/news/-/asset_publisher/cKNonkJE02Bw/content/id/2804288 (accessed date: 11.07.2024).

[93] China did not publically share its position, see: Korzak E. UN GGE on Cybersecurity: The End of an Era? – *The Diplomat.* 31 July 2017. URL: https://thediplomat.com/2017/07/un-gge-on-cybersecurity-have-china-and-russia-just-made-cyberspace-less-safe (accessed date: 11.07.2024).

[94] Declaration by Cuba, at the Final Session of Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security. New York. 23 June 2017. URL: http://misiones.minrex.gob.cu/en/un/statements/71-unga-cuba-final-session-group-governmental-experts-developments-field-information (accessed date: 15.07.2024).

[95] Ibidem.

[96] Report of the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security. UN Doc A/76/135. 14 July 2021. Para. 71(f).

[97] See, e.g.: Shereshevsky Y. 2022. International Humanitarian Law-Making and New Military Technologies. – *International Review of the Red Cross.* 2131. URL: https://international-review.icrc.org/sites/default/files/reviews-pdf/2022-11/international-humanitarian-law-making-and-new-military-technologies-920.pdf (accessed date: 11.07.2024).

[98] Carpenter C. A Better Path to a Treaty Banning "Killer Robots" Has Just Been Cleared. – *World Politics Review.* 7 January 2022. URL: www.worldpoliticsreview.com/a-better-path-to-a-treaty-banning-ai-weapons-killer-robots/ (accessed date: 12.07.2024).

[99] Deeks A. 2023. National Security AI and the Hurdles to International Regulation. – *The Digital Social Contract: A Lawfare Paper Series.* At 12-13. URL: https://ssrn.com/abstract=4405158 (accessed date: 11.07.2024).

[100] Ibid. P. 3.

This discouraging forecast does not, however, negate the necessity of discussing how existing international law applies to AI in general and to AIpCOs in particular. The interpretation of current norms and the expression of States' positions have the potential to establish a general framework for the boundaries of acceptable AI-powered activities in an informal law-making mode, even though it is likely that these discussions will mostly take place among 'like-minded' States[101], in regional unions, bilaterally, and on non-State forums in the upcoming years.

5. Conclusion

The complication of inter-State cyber operations by AI technology raises additional questions about the application of international law, in particular its norms on the use of force, to AI-powered cyber incidents. The deployment of LAWS and commitment of AIpCOs could potentially result in the initiation of a new arms race, this time involving AI, as nations seek to develop and acquire these systems in order to maintain strategic parity. This could destabilize global security and increase the risk of conflict escalation. This and other political and ethical considerations argue in favor of limiting the discretion of States in the use of AI. However, to date, the incentives for NATO States, China and Russia to agree on an international binding instrument limiting the use of AI for malicious purposes appear illusory. Based on the history of discussions on the application of international law in cyberspace and the development of rules on the responsible behaviour of States using ICTs one could suggest that corresponding discussions on AI will likely progress outside the area of developing a comprehensive treaty framework. Further analysis of this development, thus, will require examining how transnational norms, such as those emerging from soft law instruments, customary practices, and private sector initiatives, will shape the legal landscape of AI application. Understanding how international legal norms are adapted and refined in response to emerging challenges requires a dynamic analytical approach (such as, for instance, Transnational Legal Process) to address legal dialogue and interaction among different legal systems, including domestic legal systems, regional frameworks, and global norms in a continuous process of interpretation, argumentation, and practice.

References

1. Anderson K., Reisner D., Waxman M. Adapting the Law of Armed Conflict to Autonomous Weapons Systems. – *International Law Studies.* 2014. Issue 90. P. 386-411.
2. Asaro P. On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making. – *International Review of the Red Cross.* 2012. Vol. 94. Issue 886. P. 687-709.
3. Čerka P., Grigienė J., Sirbikytė G. Is It Possible to Grant Legal Personality to Artificial Intelligence Software Systems?' – *Computer Law & Security Review.* 2017. Vol. 33. Issue 5. P. 685-699.
4. Chinen M. 2023. *The International Governance of Artificial Intelligence.* Edward Elgar Publishing. 2023. 338 p.
5. Chernyavsky A.G., Sibileva O.P. Avtonomnoe vysokotochnoe oruzhie kak vyzov mezhdunarodnomu gumanitarnomu pravu [Autonomous precision weapons as a challenge international humanitarian law]. – *Military law.* 2020. № 4. P. 229-238. (In Russ.)
6. Chesterman S. Artificial Intelligence and the Limits of Legal Personality. – *International & Comparative Law Quarterly.* 2020. Vol. 69. Issue 4. P. 819-844.
7. Delerue F. *Cyber Operations and International Law.* Cambridge University Press. 2020. 549 p.
8. Finnemore M., Hollis D.B. Beyond Naming and Shaming: Accusations and International Law in Cybersecurity. – *European Journal of International Law.* 2020. Vol. 31. Issue 3. P. 969-1003.
9. Garcia D. Lethal Artificial Intelligence and Change: The Future of International Peace and Security. – *International Studies Review.* 2018. Vol. 20. Issue 2. P. 334-341.
10. Koh H.H. Transnational Legal Process. – *The Nature of International Law.* Simpson G. (ed). Routledge. 2001. P. 311-338.
11. Koh H.H. Is there a "New" New Haven School of International Law? – *Yale Journal of International Law.* 2007. Vol 32. P. 559-573.
12. Koh H.H. Transnational Legal Process and the "New" New Haven School of International Law. – *International Legal Theory. Foundations and Frontiers.* Dunoff J.L. and Pollack M.A. (eds). Cambridge University Press. 2022. P. 101-132.
13. Lee J. *Artificial Intelligence and International Law.* Springer. 2022. 261 p.
14. Morkhat P.M. Iskusstvennyj intellekt s tochki zrenija mezhdunarodnogo gumanitarnogo prava [Artificial intelligence from the perspective of international humanitarian law]. – *Law and State: Theory and Practice.* 2017. № 10. P. 18-24. (In Russ.)

[101] Particularly, the NATO states have agreed on the Principles of Responsible Use for AI. See: Stanley-Lockman Z., Christie E.H. An Artificial Intelligence Strategy for NATO. 25 October 2021. URL: www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html (accessed date: 11.07.2024).

15. Proskurina D.S., Khokhlova M.I., Safin N.I. Smertonosnye avtonomnye sistemy vooruzhenij: budushhee voennoj industrii ili ugroza padenija mezhdunarodnogo gumanitarnogo prava? [Lethal autonomous weapons systems: the future of the military industry or the threat of the fall of international humanitarian law?]. – *Eurasian Advocacy.* 2020. № 6. P. 81-85. (In Russ.)
16. Roscini M. *Cyber Operations and the Use of Force in International Law.* Oxford University Press. 2015. 307 p.
17. Roscini M. Evidentiary Issues in International Disputes Related to State Responsibility for Cyber Operations. – *Texas International Law Journal.* 2015. Vol. 50. Issue 2. P. 233-275.
18. Sassòli M. Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified. – *International Law Studies.* 2014. Vol. 90. P. 308-340.
19. Schmitt M.N. (ed.). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations.* Cambridge University Press. 2017. 598 p.
20. Schmitt M.N. Autonomous Cyber Capabilities and the International Law of Sovereignty and Intervention. – *International Law Studies.* 2020. Vol. 96. P. 549-576.
21. Stahl B.C. Responsible Computers? A Case for Ascribing Quasi-Responsibility to Computers Independent of Personhood or Agency. – *Ethics and Information Technology.* 2006. Vol. 8. Issue 4. P. 205-213.
22. Stroppa M. Legal and Ethical Implications of Autonomous Cyber Capabilities: a Call for Retaining Human Control in Cyberspace. – *Ethics and Information Technology.* 2023. Issue 25. Paper 7. P. 1-6.
23. Taddeo M., McCutcheon T., Floridi, L. Trusting Artificial Intelligence in Cybersecurity is a Double-Edged Sword. – *Nature Machine Intelligence.* 2019. Issue 1. P. 557-560.

Информация об авторе

Екатерина Александровна МАРТЫНОВА,
преподаватель департамента международного права, факультет права, Национальный исследовательский университет «Высшая школа экономики»

Мясницкая ул., д. 20, Москва, 101000, Российская Федерация

eamartynova@hse.ru
ORCID: 0000-0002-8995-4462

About the Author

Ekaterina A. MARTYNOVA,
Lecturer at the School of International Law, Faculty of Law, National Research University Higher School of Economics

20, Myasnitskaya St., Moscow, Russian Federation, 101000

eamartynova@hse.ru
ORCID: 0000-0002-8995-4462