SIPRI Insights on Peace and Security
No. 2025/03 March 2025

ADVANCING GOVERNANCE AT THE NEXUS OF ARTIFICIAL INTELLIGENCE AND NUCLEAR WEAPONS

FEI SU, VLADISLAV CHERNAVSKIKH AND WILFRED WAN*

SUMMARY

● The rapid advancement of military artificial intelligence (AI), especially its potential integration into nuclear systems, presents significant risks to strategic stability and established deterrence practices. Despite these concerns, no dedicated governance framework currently exists to address the specific challenges of the AI–nuclear nexus. Existing initiatives have primarily focused on ensuring human control over nuclear decision-making.

There are a number of state-led initiatives on the governance of military AI more broadly. They can be adapted to address the use of AI in nuclear weapons, but applying them will not be straightforward. There is thus a need to extend the conversation beyond the 'human in the loop' concept and develop targeted governance measures. Future discussions could investigate the precise level and degree of required human control and set clear red lines for both the extent and the type of AI integration in nuclear and related systems.

I. Introduction

Rapid developments in artificial intelligence (AI) capabilities have driven government investment in efforts to explore their applicability in military contexts, including in nuclear forces.[1] The extent to which AI capabilities will be adopted by the military remains debated, but they have potential utility across nuclear weapon systems. This includes in early warning and intelligence, surveillance and reconnaissance (ISR), in nuclear command, control and communications (NC3), and in delivery systems, and also in conventional systems with counterforce potential.[2] However, the use of AI in these systems would have an impact on deterrence practices and has the potential to upend strategic predictability and stability.

While reducing nuclear risks is in the collective interest of all states, there has been no discussion on establishing effective governance frameworks specifically tailored to the nexus between AI and nuclear weapons. Broader governance discussions pertaining to the use of AI in military operations have primarily focused on issues of safety, security and responsibility; these concerns are also likely to shape debates in the AI–nuclear context. However, discussions have not yet been effectively adapted to the unique challenge of nuclear forces, particularly when considering their interconnectivity with other critical technologies and domains, such as information and communications technology (ICT) and outer space.[3]

This paper assesses how current governance efforts can help to address risks associated with the integration of AI into nuclear forces. Recognizing that state-led initiatives specific to this area are limited, the paper also examines selected governance frameworks for military uses of AI in general, identifying elements relevant to governance at the AI–nuclear nexus. Section II maps current state-led governance approaches to this end and the

[1] Chernavskikh, V., 'Nuclear weapons and artificial intelligence: Technological promises and practical realities', SIPRI Background Paper, Sep. 2024.

[2] Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Rand Corp.: Santa Monica, CA, 2018) ; Boulanin, V. et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (SIPRI: Stockholm, June 2020) ; Johnson, J., *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (Oxford University Press: Oxford, 2023) ; and Saltini, A., *AI and Nuclear Command, Control and Communications: P5 Perspectives* (European Leadership Network: London, Nov. 2023).

[3] On the latter see e.g. Raju, N., 'Parameters to assess escalation risks in space', SIPRI Research Policy Paper, Feb. 2025.

* The authors would like to thank the German Federal Foreign Office, which generously provided funding for this project.

operational measures included in these initiatives. Section III identifies gaps in these existing measures and investigates challenges that may hinder progress in AI–nuclear governance. It also proposes some potential pathways forwards. The paper concludes in section IV by summarizing key findings.

II. AI governance approaches applicable to nuclear forces

This section first explores the few state-led governance initiatives that directly address the risks associated with the application of AI in nuclear weapons and related systems. These initiatives are a joint British–French–United States working paper, a US-led multinational declaration and an action plan agreed by the Summit on Responsible Artificial Intelligence in the Military Domain (REAIM). The section then looks at selected initiatives that focus on the use of AI in broader military contexts. These initiatives include a Chinese position paper, REAIM documents, the US-led declaration and a United Nations General Assembly resolution.

Governance of the use of AI in nuclear weapons

There are currently no governance frameworks that focus solely on the AI–nuclear nexus. So far, issues related to the use of AI in nuclear forces have only been discussed as elements of wider efforts in either AI governance or nuclear weapon governance. These include a working paper on principles and responsible practices for nuclear weapon states submitted in 2022 by France, the United Kingdom and the United States to the 10th Review Conference of the Treaty on the Non-Proliferation of Nuclear Weapons (NPT); a 2023 Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy issued by the USA for endorsement by other states; and, most recently, a reference to nuclear weapons in the Blueprint for Action unveiled at the 2024 REAIM Summit. While AI capabilities have accelerated significantly in recent years, approaches to the governance of their use in nuclear weapons have—and remain—centred on one critical point: that of retaining a 'human in the loop' in NC3.

Specifically, in their 2022 working paper, France, the UK and the USA committed to steps to reduce the risk of nuclear conflict, including a statement that they will 'maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment'.[4] Another point in the document, while not explicitly mentioning AI, addresses potential risks associated with the nuclear decision-making process: it highlights 'policies and procedures to ensure a deliberate process allowing leaders sufficient time to gather information and consider courses of action in a crisis'.[5] Both points suggest the need for a degree of self-restraint in potential adoption of AI in NC3, and for redundancy measures in this context to minimize the risk of accidental or inadvertent use of nuclear weapons.

[4] 10th NPT Review Conference, 'Principles and responsible practices for nuclear weapon states', Working paper submitted by France, the UK and the USA, NPT/CONF.2020/WP.70, 29 July 2022, para. 5(vii).

[5] 10th NPT Review Conference, NPT/CONF.2020/WP.70 (note 4), para. 5(v).

Building on the joint statement, in November 2023 the USA opened a political declaration for endorsement by other states.[6] The preliminary version of February 2023 had contained the same language on 'human control and involvement'; but this was removed by the time of the international launch of the declaration as US consultations with allies suggested that the AI–nuclear nexus was an area requiring 'much more discussion'.[7]

The Blueprint for Action adopted at the second REAIM Summit, in November 2024, recognized the crucial need to 'maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment'.[8] While this language replicates the text of the British–French–US working paper, REAIM provides a more inclusive platform for engagement: over 60 countries endorsed the non-binding declaration, including 4 nuclear-armed states: France, Pakistan, the UK and the USA.[9] China, which attended the summit, was not among them.

The USA has urged China and the Russian Federation to make similar commitments to maintaining human control over nuclear decision-making.[10] Progress on this front has been limited, but US President Joe Biden and Chinese President Xi Jinping 'stressed the need to maintain human control over the decision to use nuclear weapons' in a November 2024 bilateral meeting.[11] How that commitment will be further developed or operationalized remains to be seen, especially under the new US administration of President Donald J. Trump. Nevertheless, this does affirm China's position on this issue.

Governance of military uses of AI in general

Current efforts to govern the military uses of AI have overall dedicated little attention to the AI–nuclear nexus. However, many—if not most—of the measures discussed in the general debate have relevance for efforts to mitigate the risks associated with the use of AI in nuclear weapons. A survey of state-led governance initiatives of military use of AI reveals recurring themes, including accountability, capacity or capability, explainability, ethics, fairness, human control, maintenance, privacy, reliability, safety, security and transparency. These constitute key concerns that states share relating to the entire life cycle of AI technology, from design and development (including training) to deployment and ongoing performance monitoring. These concerns clearly do not just apply to potential integration of AI in the broad military context; they also have critical relevance for nuclear forces.

[6] Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, US Department of State, 9 Nov. 2023. For a list of the 58 states that had endorsed the declaration by Nov. 2024 see US Department of State, Bureau of Arms Control, Deterrence, and Stability, 'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy', 27 Nov. 2024.

[7] US Department of State, 'Political declaration on responsible military use of artificial intelligence and autonomy', 16 Feb. 2023, para. B ; and US Department of State, 'Under Secretary Jenkins' remarks at the launch event for the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy', 13 Nov. 2023.

[8] REAIM Summit, 'REAIM Blueprint for Action', 9–10 Sep. 2024, para. 5.

[9] REAIM Summit, 'Countries supporting REAIM Blueprint for Action', 9–10 Sep. 2024.

[10] Torode, G., 'US official urges China, Russia to declare only humans, not AI, control nuclear weapons', Reuters, 2 May 2024,

[11] Chinese Ministry of Foreign Affairs, 'President Xi Jinping meets with U.S. President Joe Biden in Lima', 17 Nov. 2024.

Table 1. Awareness-raising measures for the military use of artificial intelligence

Initiative	Engagement	Red lines/principles
2021 Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence	Multi-stakeholder dialogues	Non-use as a tool to start a war, pursue hegemony or undermine sovereignty and territorial security; compliance with international law
2023 REAIM Call to Action	Multi-stakeholder dialogues	Compliance with international law and relevant national, regional and international legal frameworks
2023 Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy	Continued discussions among endorsing states; engage with the rest of international community also on other relevant subjects	Compliance with international law and IHL through such measures as legal reviews
2024 REAIM Blueprint for Action	Multi-stakeholder dialogues	Compliance with international law, IHL, IHRL, other relevant legal frameworks and national law
2024 UN General Assembly Resolution 79/239, 'Artificial intelligence in the military domain and its implications for international peace and security'	Multi-stakeholder dialogues	Compliance with international law, IHL, IHRL and the UN Charter

IHL = international humanitarian law; IHRL = international human rights law; REAIM = (Summit on) Responsible Artificial Intelligence in the Military Domain; UN = United Nations.

Sources: '6th Review Conference of the Certain Conventional Weapons Convention, 'Position paper of the People's Republic of China on regulating military applications of artificial intelligence (AI)', Submitted by China, CCW/CONF.VI/WP.2, 20 Dec. 2021; REAIM Summit, 'REAIM Call to Action', 15–16 Feb. 2023; Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, US Department of State, 9 Nov. 2023; REAIM Summit, 'REAIM Blueprint for Action', 9–10 Sep. 2024; and UN General Assembly Resolution 79/239, 'Artificial intelligence in the military domain and its implications for international peace and security', 24 Dec. 2024.

The measures proposed in current military AI governance efforts can be organized in four categories (see tables 1–4): (*a*) awareness-raising measures; (*b*) responsibility measures; (*c*) safety measures; and (*d*) security measures.

Awareness-raising measures

Awareness-raising measures include principles and political statements that acknowledge the intrinsic risks posed by military uses of AI. In certain cases, these measures establish related red lines—that is, they communicate a behavioural boundary that cannot be crossed—in order to build trust and establish norms; they also include other activities that promote information-sharing and the exchange of views on relevant risks.

Notably, all military AI governance initiatives (see table 1) highlight the importance of policy dialogues and exchanges and knowledge-sharing, both globally and regionally, through official (i.e. Track 1) channels and in informal (i.e. Track 2) exchanges between academics, the private sector and other relevant non-state actors. The initiatives also specify that this engagement should be multi-stakeholder, reflecting the role of the private sector in driving many advancements in AI capabilities. Similarly, the vastness of the nuclear weapons enterprise means that it is likely that some contractors or subcontractors involved in producing relevant equipment and technology will adopt AI. This highlights the need for risk awareness throughout the entire supply chain.

In certain instances, states have suggested stances on red lines for the military use of AI. China's 2021 Position Paper on Regulating Military Applications of Artificial Intelligence—which sets out ways to strengthen ethical governance for all countries to consider and uphold—asserts that 'military applications of AI shall never be used as a tool to start a war or pursue hegemony' and opposes use of advantages in AI technology 'to undermine the sovereignty and territorial security of other countries'.[12] Moreover, there is a broad awareness of and consensus on the need for military use of AI to adhere to international humanitarian law, thereby protecting civilians and civilian objects in contexts of armed conflict. Further, the Call to Action issued by the first REAIM Summit, in 2023, emphasizes the importance of compliance with national laws and relevant regional and international legal frameworks as well as data standards.[13] Similarly, the 2024 REAIM Blueprint for Action includes a mention of the UN Charter and further stresses 'the importance of establishing national strategies, principles, standards and norms, policies and frameworks and legislations as appropriate to ensure responsible AI applications in the military domain'.[14]

When applied to the nuclear domain, red lines addressing compliance with international humanitarian law might be redundant, as the nature of nuclear weapons makes it difficult to envisage any use scenario that would not undermine sovereignty and territorial security or violate international law. For instance, the International Committee of the Red Cross (ICRC) has stated that, 'in light of their catastrophic humanitarian consequences, it is extremely doubtful that nuclear weapons could ever be used in accordance with the principles and rules of [international humanitarian law]'.[15] Experts further argue that even theoretical exceptions in which nuclear use would not violate international humanitarian law would still 'result in concrete human rights violations that are justiciable'.[16] There may, nonetheless, be utility in establishing and communicating red lines relating to the degree of integration of advanced AI in nuclear forces, for instance in ways that could encourage conflict or escalation.

It is equally important to raise awareness and build deeper understanding within current conversations, including with the public, on how the military use of AI in conventional systems might have an impact on nuclear escalation risks and, more broadly, nuclear deterrence practices.[17] For example, such risks may come from AI systems enabling advanced guidance for conventional long-range precision-strike weapons and, at the same time, improving the tracking and targeting of mobile nuclear missile launchers, thus lowering

[12] 6th Review Conference of the Certain Conventional Weapons Convention, 'Position paper of the People's Republic of China on regulating military applications of artificial intelligence (AI)', Submitted by China, CCW/CONF.VI/WP.2, 20 Dec. 2021, para. 8.

[13] REAIM Summit, 'REAIM Call to Action', 15–16 Feb. 2023.

[14] REAIM Summit (note 8), para. 8.

[15] United Nations, General Assembly, First Committee, Statement by the International Committee of the Red Cross (ICRC), 11 Oct. 2023, p. 1.

[16] International Law and Policy Institute (ILPI) and Geneva Academy of International Humanitarian Law and Human Rights, *Nuclear Weapons Under International Law: An Overview* (ILPI/Geneva Academy: Geneva, Oct. 2014), p. 7.

[17] Boulanin et al. (note 2), pp. 27–30.

Table 2. Responsibility measures for the military use of artificial intelligence

Initiative	Human control and accountability	Capacity-building
2021 Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence	Accountability mechanism	Training for operators; international cooperation to support developing countries
2023 REAIM Call to Action	Humans remain responsible and accountable; ensure human oversight	Training for operators; knowledge-sharing between states and among multi-stakeholders
2023 Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy	Senior official oversight over the development and deployment of AI capabilities	Training for both the operator and those who approve the use of military AI
2024 REAIM Blueprint for Action	Human control over decisions concerning nuclear weapon deployment; human-centric, responsible and accountable use of AI in the military domain; human judgment and control over the use of force	Training programmes for military personnel; support developing countries; international cooperation
2024 UN General Assembly Resolution 79/239, 'Artificial intelligence in the military domain and its implications for international peace and security'	Human judgment and control over the use of force	Knowledge-sharing and especially support developing countries

IHL = international humanitarian law; IHRL = international human rights law; REAIM = (Summit on) Responsible Artificial Intelligence in the Military Domain; UN = United Nations.

Sources: '6th Review Conference of the Certain Conventional Weapons Convention, 'Position paper of the People's Republic of China on regulating military applications of artificial intelligence (AI)', Submitted by China, CCW/CONF.VI/WP.2, 20 Dec. 2021; REAIM Summit, 'REAIM Call to Action', 15–16 Feb. 2023; Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, US Department of State, 9 Nov. 2023; REAIM Summit, 'REAIM Blueprint for Action', 9–10 Sep. 2024; and UN General Assembly Resolution 79/239, 'Artificial intelligence in the military domain and its implications for international peace and security', 24 Dec. 2024.

the survivability of an adversary's nuclear deterrent forces and incentivizing nuclear use in a crisis.[18]

Responsibility measures

Responsibility measures emphasize human responsibility and liability in the development and use of AI for military operations. They also involve capacity-building and training programmes with a view to enhancing operators' understanding of the technologies they use—in terms of both technical literacy and awareness of legal and ethical issues. In all the examined initiatives, there is a consistent emphasis on human control, responsibility and accountability in the use of AI for military purposes (see table 2). Notably, China's 2021 Position Paper proposes that each state establish an accountability mechanism for ensuring human responsibility.[19] The US-led Political Declaration highlights the role of senior officials in overseeing

[18] Stokes, J. et. al., *Averting AI Armageddon: U.S.–China–Russia Rivalry at the Nexus of Nuclear Weapons and Artificial Intelligence* (Center for a New American Security: Washington, DC, Feb. 2025), pp. 3–6.

[19] 6th Review Conference of the Certain Conventional Weapons Convention, CCW/CONF.VI/WP.2 (note 12), para. 11.

‘the development and deployment of military AI capabilities with high-consequence applications’.[20]

As mentioned above, the importance of human control has emerged as a foundational element in discussions surrounding AI applications in nuclear weapons. Yet questions remain. For instance, the text of the 2024 REAIM Blueprint for Action leaves open the interpretation of what ‘human control and involvement’ entails. Further complicating factors are the nuclear-armed states’ differing national models for authorizing the use of nuclear weapons and their varying levels of transparency with regards to the nuclear command chain.[21] These challenge efforts to define what maintaining human control would mean in practice. Furthermore, even with oversight and control, there are risks of human operators either overly relying on the output of an AI system or, conversely, overly mistrusting it. If this output relates, for instance, to the nature of a detected incoming attack, potential repercussions could include a nuclear response.[22]

In discussions about military AI governance, there is a strong emphasis on implementing structured and continuous training programmes for the personnel operating AI-enabled systems. Such programmes are designed to provide individuals with the expertise necessary to understand the systems’ capabilities, limitations and potential consequences. The US-led Political Declaration further highlights the importance of equipping personnel who use or approve the use of military AI capabilities with the knowledge and skills needed to exercise sound judgment and make informed decisions.[23]

There are also external aspects of capacity-building, with the initiatives seeking to promote international knowledge-sharing frameworks and to provide support to developing countries in order to address disparities in understanding the risks and challenges associated with AI integration in the military domain.[24] Notably, these engagement and capacity-building activities can also be seen as a way to shape future norms on AI. At the launch of the 2023 Political Declaration, US Vice President Kamala Harris noted the need ‘to lay the groundwork for the future of AI’ and ‘to create a collective vision of what this future must be’.[25] These efforts to foster inclusive participation can align with broader efforts in the nuclear sphere to increase understanding of the risk of misinterpretation and miscalculation, including by involving officials from both nuclear-armed and non-nuclear-armed states, as outlined in the British–French–US working paper.[26]

[20] Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (note 6), para. C.

[21] Lewis, J. G. and Tertrais, B., *The Finger on the Button: The Authority to Use Nuclear Weapons in Nuclear-armed States*, James Martin Center for Nonproliferation Studies (CNS) Occasional Paper no. 45 (CNS: Monterey, CA, Feb. 2019).

[22] Depp, M. and Scharre, P., ‘Artificial intelligence and nuclear stability’, War on the Rocks, 16 Jan. 2024.

[23] Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (note 6), para. G.

[24] 6th Review Conference of the Certain Conventional Weapons Convention, CCW/CONF.VI/WP.2 (note 12), para. 14; REAIM Summit (note 8), para. 15; and UN General Assembly Resolution 79/239, ‘Artificial intelligence in the military domain and its implications for international peace and security’, 24 Dec. 2024, p. 2.

[25] US Department of State, ‘Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy’, Media note, 13 Nov. 2023.

[26] 10th NPT Review Conference, NPT/CONF.2020/WP.70 (note 4), para. 5.

Table 3. Safety measures for the military use of artificial intelligence

Initiative	Data governance	Maintenance and monitoring
2021 Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence	Restricting military use of AI data	–
2023 REAIM Call to Action	Data-protection and data-quality governance mechanisms	–
2023 Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy	Transparency and auditability of data sources	Continuous monitoring and testing of AI systems throughout their life cycle
2024 REAIM Blueprint for Action	Data-governance mechanisms, including clear policies and procedure	Continuous monitoring processes
2024 UN General Assembly Resolution 79/239, 'Artificial intelligence in the military domain and its implications for international peace and security'	–	–

IHL = international humanitarian law; IHRL = international human rights law; REAIM = (Summit on) Responsible Artificial Intelligence in the Military Domain; UN = United Nations.

Sources: '6th Review Conference of the Certain Conventional Weapons Convention, 'Position paper of the People's Republic of China on regulating military applications of artificial intelligence (AI)', Submitted by China, CCW/CONF.VI/WP.2, 20 Dec. 2021; REAIM Summit, 'REAIM Call to Action', 15–16 Feb. 2023; Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, US Department of State, 9 Nov. 2023; REAIM Summit, 'REAIM Blueprint for Action', 9–10 Sep. 2024; and UN General Assembly Resolution 79/239, 'Artificial intelligence in the military domain and its implications for international peace and security', 24 Dec. 2024.

Safety measures

Safety measures aim to ensure reliable performance of AI systems in order to prevent accidents. The military AI governance initiatives demonstrate common concerns over the risk of malfunction and accompanying consequences, with calls for safeguards to mitigate potential system failures. In particular, the 2023 US-led Political Declaration emphasizes the need for continuous monitoring and testing of AI systems to ensure that they operate as intended and that critical safety features remain intact throughout their life cycle.[27] The 2024 REAIM Blueprint for Action calls for greater engagement in developing testing and evaluation protocols as well as robust monitoring processes.[28]

The approach taken by states to civilian nuclear safety issues—which centre on safety standards established by the International Atomic Energy Agency (IAEA) as well as peer review and other technical cooperation mechanisms—constitutes a potential path forwards.[29] In the case of nuclear weapons, however, NPT non-proliferation obligations limit the ability of states to work multilaterally to advance these processes. Additionally, non-proliferation compliance has previously been primarily monitored through observable physical activities and the gathering of tangible evidence—but these methods

[27] Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (note 6), para. I.

[28] REAIM Summit (note 8), para. 10.

[29] E.g. International Atomic Energy Agency (IAEA), *IAEA Safeguards: Serving Nuclear Non-Proliferation* (IAEA: Vienna, 2024).

are less compatible with AI-related protocols.[30] Yet the work of initiatives that involve both nuclear-armed and non-nuclear weapon states may be instructive for building confidence in compliance in a context where there are barriers to information-sharing. One example of this type of initiative is the International Partnership on Nuclear Disarmament Verification, which was launched by the USA in 2014 and brings together representatives from 30 countries and the European Union.[31]

Moreover, robust data governance is essential for ensuring the reliability of AI systems. A key component of the examined initiatives is the establishment of data-governance frameworks aimed at protecting and maintaining the quality of data and standardizing practices across the AI data life cycle—from collection to deletion. Specifically, the 2023 US-led Political Declaration underlines the necessity of the sources of data used in AI systems being transparent and auditable to the relevant military personnel.[32] The REAIM Call to Action and Blueprint for Action both highlight the critical need for mechanisms to govern data quality and the protection of data.[33] Most such data-oriented efforts are likely to remain at the national level given military sensitivities and the desire to retain competitive advantage. However, there may be exceptions for alliance relationships. For instance, pillar II of the trilateral security partnership between Australia, the UK and the USA (AUKUS) specifically seeks to enhance interoperability and data exchange, including in AI, with early work on deploying 'common advanced [AI] algorithms on multiple systems'.[34] The ultimate aim of these initiatives is to allow for any combinations of data sets, AI models, algorithms and platforms to be used reliably across the militaries of all three states.[35] In nuclear contexts, however, procedures around data auditability would necessarily be internal, again given non-proliferation obligations under the NPT, or at most confined to selected nuclear weapon states (e.g. through the 1958 British–US Mutual Defense Agreement).[36]

Security measures

Security measures seek to protect the secure performance of AI systems and reduce vulnerabilities that may expose systems to adversarial attacks, or to prevent AI systems themselves from falling into the wrong hands. Integration of AI into military contexts may bring with it increased cybersecurity risks, including those related to adversarial actions aimed at compromising

[30] Stewart, I. J., 'Why the IAEA model may not be best for regulating artificial intelligence', *Bulletin of the Atomic Scientists*, 9 June 2023.

[31] International Partnership for Nuclear Disarmament Verification (IPNDV), *Verification of Nuclear Disarmament: Insights from a Decade of the International Partnership for Nuclear Disarmament Verification* (Nuclear Threat Initiative: Washington, DC, June 2024).

[32] Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (note 6), para. F.

[33] REAIM Summit, 'REAIM Call to Action' (note 13), para. 10; and REAIM Summit, 'REAIM Blueprint for Action' (note 8), para. 17.

[34] AUKUS Defence Ministers Meeting, Joint statement, 1 Dec. 2023.

[35] US Department of Defense, 'AUKUS pillar II milestones hint at future integrated autonomous, artificial intelligence operations', News release, 9 Aug. 2024.

[36] British–US Agreement for Cooperation on the Uses of Atomic Energy for Mutual Defense Purposes, signed 3 July 1958, entered into force 4 Aug. 1958, Treaty Series no. 41, Oct. 1958.

Table 4. Security measures for the military use of artificial intelligence

Initiative	Proliferation	Adversarial attack
2021 Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence	Mitigating proliferation risks of military applications of AI	–
2023 REAIM Call to Action	–	–
2023 Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy	–	–
2024 REAIM Blueprint for Action	Prevent AI technologies from being used to contribute to WMD proliferation by states and non-state actors including terrorist groups	Cybersecurity: AI in cyber operations, AI in electronic warfare and AI in information operations
2024 UN General Assembly Resolution 79/239, 'Artificial intelligence in the military domain and its implications for international peace and security'	Addressing proliferation to non-state actors	–

IHL = international humanitarian law; IHRL = international human rights law; REAIM = (Summit on) Responsible Artificial Intelligence in the Military Domain; UN = United Nations; WMD = Weapons of mass destruction.

Sources: '6th Review Conference of the Certain Conventional Weapons Convention, 'Position paper of the People's Republic of China on regulating military applications of artificial intelligence (AI)', Submitted by China, CCW/CONF.VI/WP.2, 20 Dec. 2021; REAIM Summit, 'REAIM Call to Action', 15–16 Feb. 2023; Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, US Department of State, 9 Nov. 2023; REAIM Summit, 'REAIM Blueprint for Action', 9–10 Sep. 2024; and UN General Assembly Resolution 79/239, 'Artificial intelligence in the military domain and its implications for international peace and security', 24 Dec. 2024.

training data or real-time data fed to AI systems.[37] Notably, however, most multilateral initiatives on military AI governance (with the exception of the 2024 REAIM Blueprint for Action) make only limited explicit mentions of the threat of adversarial attacks.[38] Rather, the threat is often implied through an emphasis on limiting access to AI data—and it is often introduced in the context of data protection.[39] Explicit mention of concerns about vulnerability centres largely on the principle of preventing the proliferation of AI technologies to irresponsible actors, including both state and non-state actors such as terrorist groups. This concern is highlighted in China's 2021 Position Paper and the 2024 REAIM Blueprint for Action, as well as the 2024 UN General Assembly resolution on 'Artificial intelligence in the military domain and its implications for international peace and security', which was based on the outcome of the 2024 REAIM Summit.[40]

Similar concerns formed the foundation for UN Security Council Resolution 1540 of 2004.[41] This obliges UN member states to refrain from providing support to non-state actors that seek to manufacture, acquire, possess, develop, transport, transfer or use nuclear, chemical or biological weapons and their means of delivery, and to adopt and enforce legislation

[37] Saltini, A., 'Navigating cyber vulnerabilities in AI-enabled military systems', European Leadership Network, 19 Mar. 2024.

[38] REAIM Summit (note 8), para. 4.

[39] E.g. 6th Review Conference of the Certain Conventional Weapons Convention, CCW/CONF.VI/WP.2 (note 12), para. 10.

[40] UN General Assembly Resolution 79/239 (note 24).

[41] UN Security Council Resolution 1540, 28 Apr. 2004.

to prohibit these activities. Similarly, export control regimes—including the Nuclear Suppliers Group (NSG) and the Zangger Committee—have been key in preventing the supply of goods and technologies that could contribute to nuclear weapon programmes. These frameworks also cover a range of hardware, software and technology that enables or is related to dual-use AI.[42] However, the enabling nature and potential intangibility of AI, as well as the dual-use nature of both AI and nuclear technologies, can compound the difficulties of addressing proliferation issues at the AI–nuclear nexus through export controls.

III. Challenges and pathways to governance of AI in nuclear forces

The above overview of military AI governance initiatives—including their approaches and proposed measures—provides indications as to their applicability to nuclear forces. In some cases, there is clear overlap, suggesting potential for complementarity and for the elaboration of foundational principles for the regulation of the AI–nuclear nexus. At the same time, the limitations of this overlap, as well as the limitations of general efforts to govern military AI, also become apparent. This section examines the challenges to governance posed by AI and nuclear weapons—both individually and at their intersection—and explores potential paths forwards. The discussion is structured around three types of challenge that emerge: conceptual challenges; political and institutional challenges; and implementation challenges.

Conceptual challenges

As a general-purpose technology with a wide range of use cases and applications, AI presents a fundamental challenge to the design of multilateral governance measures.[43] The complex and ambiguous nature of AI as a technological field leads to definitional uncertainty and inconsistent interpretations among states that attempt to address its use in military contexts.[44] The absence of a commonly agreed vocabulary can result in a regulatory and governance landscape with varying boundaries for applicability. This is especially relevant for advanced AI, where the relationship between the technical parameters of a model and its actual capabilities is becoming increasingly complex.[45] The very notion of 'military AI' is also evolving, as military, intelligence and diplomatic communities all make use of different AI applications to carry out their specific missions in order to

[42] Boulanin, V., Brockmann, K. and Richards, L., *Responsible Artificial Intelligence Research and Innovation for International Peace and Security* (SIPRI: Stockholm, Nov. 2020) ; and Héau, L. and Brockmann, K., *Intangible Transfers of Technology and Software: Challenges for the Missile Technology Control Regime* (SIPRI: Stockholm, Apr. 2024).

[43] Boulanin, V., 'Regulating military AI will be difficult. Here's a way forward', *Bulletin of the Atomic Scientists*, 3 Mar. 2021.

[44] Boulanin, V., 'Artificial intelligence: A primer', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019).

[45] Gupta, R. et al., 'Data-centric AI governance: Addressing the limitations of model-focused policies', arXiv 2409.17216, 25 Sep. 2024, pp. 2–6.

gain strategic or battlefield advantage.[46] Furthermore, the rapid pace of technological advances—and the increased efficiency of AI models—suggest that regulatory ceilings defined by technical characteristics could quickly become anachronistic.[47]

Governance is further complicated because advanced AI models are defined by their generalized capabilities and their potential to be relatively easily adapted to a variety of military tasks.[48] The same data, algorithms and computational power can be adapted for both civilian and military objectives. For example, in 2024, the Chinese People's Liberation Army (PLA) built an AI tool tailored for military intelligence analysis and decision-making support by adapting a publicly available large language model developed by the US company Meta.[49] It is difficult to simply regulate all potential downstream applications of these models as laws, norms or policies that are adequate for one application may be inapplicable or inappropriate for another. Indeed, even if significant progress is made on governance of AI in military operations, it may not be relevant for nuclear forces; and then any agreement on general principles on integration of AI into NC3 may not be applicable to early-warning, ISR, delivery or other nuclear systems.

All of this suggests that regulating the AI–nuclear nexus may require a two-tier approach in which general norms and principles of behaviour are sustained by a steady stream of adaptable and system-specific measures. Using operational definitions and focusing on nuclear weapons at both stages may help in navigating some of the issues listed above. As a first step, a general principle in the form of a commitment to maintain human control over nuclear decision-making could be expanded to all nuclear-armed states. Additionally, a concerted effort could be made by nuclear-armed states to identify and communicate the nuclear-related systems where potential integration of advanced AI capabilities raises greatest concerns. This effort could take the form of a sustained dialogue involving nuclear-armed states or voluntary information-sharing arrangements. It would be undertaken with a view to establishing, and committing to, additional technical red lines centred on AI integration in particular NC3 systems or nuclear-delivery systems.[50] These red lines could be complemented by coordinated national regulations and norms concerning the design and use of military AI systems—in such contexts as air and missile defence complexes or underwater and space ISR assets—that have an impact on nuclear weapons.

Such an approach would help disentangle the conversation on governance of the AI–nuclear nexus from definitional and conceptual ambiguities that are characteristic of the AI field. It would also help to ensure that any adopted regulation measures remain relevant regardless of how AI technology advances. Proposals regarding a similar two-tier approach to regulating autonomous weapon systems (AWS) have gained popularity in

[46] Rosen, B., 'How to make military AI governance more robust', War on the Rocks, 6 Aug. 2024.

[47] Rosen (note 46).

[48] Hickey, A., 'The GPT dilemma: Foundation models and the shadow of dual-use', arXiv 2407.20442, 29 July 2024.

[49] Cheung, S., 'PRC adapts Meta's Llama for military and security AI applications', China Brief, Jamestown Foundation, 31 Oct. 2024.

[50] Johnson, J., *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (Oxford University Press: Oxford, 2023), p. 194 ; and Lamberth, M. and Scharre, P., 'Arms control for artificial intelligence', *Texas National Security Review*, vol. 6, no. 2 (spring 2023).

UN discussions in recent years. In this context, they suggest, first, a complete prohibition on certain types and uses of AWS and, second, imposing limits and requirements on the development and use of all AWS that are not prohibited.[51]

Political and institutional challenges

Another slew of challenges comes from the political and institutional environments in which nuclear weapons function. First, AI integration can be seen by nuclear-armed states as an opportunity to bolster nuclear capabilities. This may dissuade them from early engagement on risk mitigation if they consider that it will be detrimental to this goal. Second, sensitivity inherently associated with military nuclear programmes and political strife in international nuclear governance frameworks make multi-stakeholder dialogue on issues stemming from the AI–nuclear nexus difficult. Finally, the most appropriate and effective forum to address the AI–nuclear nexus is yet to be determined. These three sets of challenges are explored in more detail below.

Potential rewards of integration

Discussion on risks of nuclear escalation and measures to reduce them features prominently in multilateral nuclear forums. However, the full implications of this conversation for AI integration in nuclear forces are not yet clear to all parties. Given the current strategic context, the prevailing approach among nuclear-armed states and their allies is to strengthen deterrence—including through extensive nuclear modernization programmes, and potentially through AI–nuclear integration.[52] States that perceive themselves as being at a strategic disadvantage may be willing to accept the risks that are associated with AI integration.[53] Russia was reportedly reluctant to engage with US attempts to negotiate a shared commitment to maintain human control over nuclear use decisions among the five NPT-recognized nuclear weapon states—China, France, the Russian Federation, the United Kingdom and the United States (the P5).[54] Concerns have also emerged that the second US administration of President Trump could intensify military use of AI in pursuit of a strategic advantage, potentially reversing some of the 'responsible' approaches formulated under the previous, 2021–25 administration of President Joe Biden.[55] However, it is worth noting that the first Trump administration, of 2017–21, established major policy frameworks concerning the 'ethical' use of AI in defence. Additionally, the continuity

[51] Bruun, L., *Towards a Two-tiered Approach to Regulation of Autonomous Weapon Systems: Identifying Pathways and Possible Elements* (SIPRI: Stockholm, Aug. 2024).

[52] On these programmes see 'World nuclear forces', *SIPRI Yearbook 2024: Armaments, Disarmament and International Security* (Oxford University Press: Oxford, 2024).

[53] Saltini, A. and Pan, Y., 'Beyond human-in-the-loop: Managing AI risks in nuclear command-and-control', War on the Rocks, 6 Dec. 2024.

[54] 2026 NPT Review Conference, Preparatory Committee, Statement by the USA, 22 July 2024.

[55] Zakrzewski, C., 'Trump allies draft AI order to launch "Manhattan Projects" for defense', *Washington Post*, 16 July 2024 ; and Saltini, A., 'AI and nuclear strategy under Trump 2.0: What to expect', Open Nuclear Network, 30 Jan. 2025.

of internal institutional processes within US national security institutions might mitigate radical policy shifts.[56]

It is revealing that the foundational principle at the AI–nuclear nexus has been the 'human in the loop'. This narrowly addresses human control over nuclear use decisions, and thus primarily addresses risk of inadvertent nuclear weapon use. This aligns with the focus in broader nuclear risk-reduction conversations on measures that would lower the possibility of misperception, miscalculation and misunderstanding without requiring fundamental change to states' nuclear behaviours, capabilities or strategies. This then suggests that multilateral regulation at the AI–nuclear nexus in the short-to-medium term will have to fall along these lines in order to be politically acceptable for nuclear-armed states.

A prerequisite for the further elaboration of general principles or specific measures is the identification of scenarios of inadvertent escalation linked to the integration of AI into nuclear weapons and conventional systems. For instance, several sessions at the 2024 REAIM Summit emphasized the need to focus on strategic security as well as testing and analysis of AI use in concrete case studies.[57] Given the high stakes involved, any integration in the military domain that has implications for nuclear decision-making should constitute a natural part of this conversation.

Nuclear weapon politics

While REAIM and other military AI governance initiatives have sought to involve a diverse group of stakeholders—including non-nuclear-armed states, expert communities and the private sector—the nature of nuclear weapons can hinder such approaches. Engagement from expert communities and the private sector, which is a feature of general AI governance conversations, will be less likely in the case of AI–nuclear integration, even if the private sector is the most responsible for technological advancements in AI. This is because nuclear weapon programmes remain one of the most sensitive areas even within domestic military and government structures, with knowledge and access confined to a privileged few. This suggests that the development of governance efforts—especially those linked to responsibility and safety measures, including relating to data standards—will require the establishment of a significant level of trust between states before protocols can be discussed, let alone data. Establishment of trust could be facilitated if the involved states establish a verification regime designed to prevent breaches of secrecy; but this has its own set of challenges (as explored below). Nevertheless, the private sector and expert communities can offer input on major risks associated with AI integration and can provide technical expertise outside the most sensitive elements of nuclear weapon complexes.

The different 'classes' of states and related politics in the nuclear landscape can present additional barriers to multilateral AI–nuclear governance. The five nuclear weapon states recognized under the NPT (also the P5), nuclear-armed states outside the NPT, and non-nuclear weapon states involved

[56] Rosen, B., 'The AI presidency: What "America first" means for global AI governance', Just Security, 16 Dec. 2024.

[57] Rosen, B., 'From principles to action: Charting a path for military AI governance', Carnegie Council for Ethics in International Affairs, 12 Sep. 2024.

in nuclear alliances—these groups often hold fundamentally different perspectives on issues within the nuclear domain.

Some states have raised concerns in the ongoing NPT review cycle about AI potentially multiplying nuclear risks.[58] Moreover, at the initiative of the USA, the use of AI in NC3 systems has also been discussed among the P5.[59] Yet these discussions exclude the four nuclear-armed states outside the NPT (i.e. India, Israel, North Korea and Pakistan), which can limit not only the reach of regulatory approaches but also their substance, as any measures agreed in an NPT forum might be perceived by the P5 as putting themselves at a strategic disadvantage against the other four states. AI–nuclear discussions suffer from political roadblocks linked to those processes as NPT states parties have failed to reach consensus at regular review conferences since 2010 and progress in the P5 process has stalled since Russia's full-scale invasion of Ukraine. In a contentious geopolitical environment, AI integration issues are already becoming linked to other entrenched nuclear policy disagreements.[60]

In this regard, substantive informal engagement in Track 1.5 (i.e. mixed official and informal) and Track 2 dialogues (informal) may be needed before the discussion can occur on the official, Track 1 level.

Identifying the right forum

Further mainstreaming the AI–nuclear nexus into discussions on governance of military AI at such forums as REAIM could present a more productive means of making progress. Insights generated from these discussions could then be integrated into broader forums or taken forwards within UN forums, including the General Assembly's First Committee (on disarmament and international security).[61] Yet, as suggested above, the relatively slow pace and limited progress of discussions on the governance of military AI, notably AWS, combined with the inclusive approach to discussions might make those forums less effective for addressing issues linked to AI integration in nuclear forces. Furthermore, a significant challenge to operationalizing measures to address the AI–nuclear nexus through the UN disarmament machinery is the fact that discussions focusing on risk reduction are perceived by some non-nuclear weapon states as attempts to further normalize the possession of nuclear weapons and as detrimental to the disarmament process.[62] This suggests the need for further groundwork and confidence-building, including outside the rubric of 'nuclear risk reduction' and beyond traditional disarmament forums.

Accordingly, some experts have suggested creating dedicated spaces for discussions on the AI–nuclear nexus, both bilateral (e.g. between China and the USA) and in ad hoc multilateral frameworks.[63] One example of an ad hoc

[58] 2026 NPT Review Conference, Preparatory Committee, Chair's summary, NPT/CONF.2026/PC.II/WP.44, 5 Aug. 2024.

[59] 2026 NPT Review Conference, Statement by the USA (note 54).

[60] Rand, L., 'The risk of bringing AI discussions into high-level nuclear dialogues', Carnegie Endowment for International Peace, 19 Aug. 2024.

[61] Unal, B. and Richard, U., *Governance of Artificial Intelligence in the Military Domain*, UN Office for Disarmament Affairs (UNODA) Occasional Papers no. 42 (United Nations: New York, June 2024).

[62] 2026 NPT Review Conference, Preparatory Committee, Statement by the New Agenda Coalition, 24 July 2024.

[63] Saltini (note 2); and Renshaw, J. and Hunnicutt, T., 'Biden, Xi agree that humans, not AI, should control nuclear arms', Reuters, 17 Nov. 2024.

multilateral forum where AI issues have recently emerged is the Creating an Environment for Nuclear Disarmament (CEND) initiative launched by the USA, in which officials from over 40 states participate, including nuclear weapon states recognized under the NPT, nuclear-armed states outside the NPT and non-nuclear-armed states.[64] While the diverse perspectives represented and its informal working methods mean that CEND is unlikely to reach a consensus, it may facilitate the kind of frank exchange on risks and escalation pathways that is necessary to advance the conversation on governance at the AI–nuclear nexus. To complement this, the experience of UN groups of governmental experts and open-ended working groups in elaborating general principles in other technological fields (e.g. AWS, cyberspace and outer space) in recent years could present technically grounded approaches with which to engage on specific relevant use cases of AI relevant to nuclear weapons.[65] Ultimately, creating a standalone platform for engagement on the AI-nuclear nexus could be a valuable step forwards as such a forum could help bridge existing conversations and processes on technological advancements relevant to international security in other domains, the military applications of AI and nuclear weapon issues.

Implementation challenges

Even if states were to reach an agreement on specific governance measures to address the AI–nuclear nexus, implementation would still pose significant challenges. When addressing the deployment of AI in military systems, it is extremely difficult to verify when and whether AI tools have been used as they do not necessarily alter the physical signature of weapon systems. It may be even harder to determine the extent to which AI influenced the behaviour of specific systems, the stages at which it had an impact on decision-making and the role of human involvement.[66] When applied to the sensitive area of nuclear weapons and related systems, these challenges are especially pertinent since the level of access and transparency that is necessary for verification is practically impossible, even in cooperative regimes.

At the AI–nuclear nexus then, the most feasible way forwards may be to employ a variety of risk-mitigation and confidence-building measures that could help develop common sets of norms and practices, with a view to their eventually being formally codified into political commitments and multilateral agreements. Among like-minded states, including in the context of cooperative security agreements such as AUKUS, it may be possible to develop consensus on standards and procedures to evaluate AI capabilities for specific use cases, as well as a framework for risk assessment of AI integration. These could then become frameworks applicable to nuclear weapons or could even facilitate the negotiation of a document or statement—between

[64] US Department of State, Bureau of Arms Control, Deterrence, and Stability, 'Creating an Environment for Nuclear Disarmament (CEND)', [n.d.].

[65] Rand (note 60). On recent developments in these processes see e.g. 'International governance of artificial intelligence, cyberspace and outer space', *SIPRI Yearbook 2024* (note 52).

[66] Sweijs, T. and Romansky, S., *International Norms Development and AI in the Military Domain*, Centre for International Governance Innovation (CIGI) Papers no. 300 (CIGI: Waterloo, ON, Sep. 2024) ; and Sastry, G. et al., 'Computing power and the governance of artificial intelligence', arXiv 2402.08797, 14 Feb. 2024.

the states involved and potentially universalized among all nuclear-armed states—on principles for responsible use of AI in nuclear forces.[67]

Additional areas for potential cooperation relevant to the AI–nuclear nexus involve initiatives centred on testing, evaluation, validation and verification (TEVV) practices. Like-minded states could initially establish national standards for TEVV protocols for AI in military applications, focusing on applications with relevance for nuclear decision-making, and could standardize methodologies for red teaming (i.e. simulated attacks) for such applications.[68]

Longer-term risk-reduction measures could be centred on the exchange of information on results of the evaluation and red teaming of advanced AI models. This could be accompanied by the establishment of international hotlines; joint data centres; risk-reduction centres on AI incidents; and military-to-military dialogues on AI in order to exchange on doctrines and rules of engagement.[69] In data governance, like-minded states could first seek to establish a platform for multi-stakeholder coordination, with the involvement of industry and civil society experts, in the context of military applications of AI technologies. This would allow development of common regulatory approaches to curating training data sets and standardizing data set documentation.[70] To address safety and security concerns, states might seek to develop norms on avoiding the data poisoning of safety-critical data streams and to cooperate on detection of data poisoning.[71]

Given the widespread and increasingly accessible nature of AI technology, a potential area for initial collaboration could involve developing measures to safeguard AI-enabled systems from malicious interference by non-state actors. All of the above would strengthen the broader regulatory framework in which AI–nuclear integration would take place.

IV. Conclusions

Despite the rapid deployment of military AI capabilities and their increasing use on battlefields, progress in governing military AI remains limited or absent—particularly in discussions surrounding its application in nuclear weapons. However, existing approaches to the governance of military AI reveal common areas of concern that can be addressed through coordinated measures. Military AI governance initiatives and proposed measures aim to address four key shared concerns: (*a*) raising awareness through multi-stakeholder engagement and clarifying red lines and compliance with international laws; (*b*) ensuring responsibility and preserving the capacity for meaningful human intervention; (*c*) implementing safety measures to address data security and ensuring the reliable performance of AI systems;

[67] Wehsener, A. et al., *AI–NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures* (Institute for Security and Technology: Oakland, CA, Feb. 2023) ; and Saltini (note 2).

[68] Horowitz, M. and Scharre P., *AI and International Stability: Risks and Confidence-Building Measures* (Center for a New American Security: Washington, DC, Jan. 2021).

[69] Puscas, I., *Confidence-building Measures for Artificial Intelligence A Multilateral Perspective* (UNIDIR: Geneva, 2024) ; and Wehsener et al. (note 67), p. 35.

[70] Afina, Y. and Persi Paoli, G., *Governance of Artificial Intelligence in the Military Domain: A Multi-stakeholder Perspective on Priority Areas* (UNIDIR: Geneva, 2024).

[71] Wehsener et al. (note 67).

and (*d*) applying security measures to mitigate potential threats from external sources. All of these have relevance for nuclear forces, although there can be further elaboration of measures for that specific context, some of which are highlighted above.

Meanwhile, there persist significant political and technical challenges to advancing governance of the use of AI in nuclear forces. Regulatory frameworks struggle to keep pace with AI advancements, and verification challenges remain central, particularly regarding how AI is being integrated into nuclear weapon systems. While it is crucial to involve non-nuclear-armed states, expert communities and the private sector, their limited knowledge of nuclear force structures may constrain their contributions. Fundamentally, the complex geopolitical landscape poses additional barriers to achieving a multilateral and multi-stakeholder approach. Restrictions on AI integration could be perceived by nuclear-armed states as potentially undermining their deterrence capabilities. This could then discourage efforts to move beyond commitments centred on maintaining a ‘human in the loop’.

Nonetheless, maintaining the appropriate level of human control remains a central concern at the AI–nuclear nexus. With growing consensus on the importance of human oversight, future discussions could shift towards specifying the precise level and degree of human control required in AI systems. In particular, by limiting AI’s role in the broader nuclear decision-making environment, this shift would provide an opportunity to establish and communicate red lines for both the extent and the type of integration of AI in nuclear systems to ensure that it does not undermine stability or create escalatory risks. A key focus of these discussions could involve identifying and developing sets of comprehensive technical parameters and norms to address specific challenges at the AI–nuclear nexus.

Abbreviations

AI	Artificial intelligence
AUKUS	Trilateral security partnership between Australia, the United Kingdom and the United States
AWS	Autonomous weapon system
CEND	Creating an Environment for Nuclear Disarmament
ISR	Intelligence, surveillance and reconnaissance
NC3	Nuclear command, control and communications
NPT	Treaty on the Non-Proliferation of Nuclear Weapons (Non-Proliferation Treaty)
P5	The five NPT-recognized nuclear weapon states (China, France, the Russian Federation, the United Kingdom and the United States)
REAIM	(Summit on) Responsible Artificial Intelligence in the Military Domain
TEVV	Testing, evaluation, validation and verification
UN	United Nations

SIPRI INSIGHTS ON PEACE AND SECURITY NO. 2025/03

ADVANCING GOVERNANCE AT THE NEXUS OF ARTIFICIAL INTELLIGENCE AND NUCLEAR WEAPONS

FEI SU, VLADISLAV CHERNAVSKIKH AND WILFRED WAN

CONTENTS

ABOUT THE AUTHORS

Fei Su is a researcher with the SIPRI China and Asia Security Programme. Her research interests focus on regional security issues in East Asia with a special interest in North Korea, China's foreign and security policy, maritime security, cybersecurity and technology policies.

Vladislav Chernavskikh is a research assistant with the SIPRI Weapons of Mass Destruction Programme, focusing on nuclear disarmament and non-proliferation issues. He is also a contributor to SIPRI-led activities under the Alva Myrdal Centre working group on nuclear disarmament in policy and international law.

Dr Wilfred Wan is the director of the SIPRI Weapons of Mass Destruction Programme. His recent research has focused on nuclear weapon risk reduction, nuclear disarmament verification and other issues related to arms control and disarmament. He also serves as the co-chair of the Alva Myrdal Centre working group on nuclear disarmament in policy and international law.

sipri

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE
Signalistgatan 9
SE-169 72 Solna, Sweden
Telephone: +46 8 655 97 00
Email: sipri@sipri.org
Internet: www.sipri.org

DOI: https://doi.org/10.55163/ONLB7070