SIPRI Background Paper
September 2024

NUCLEAR WEAPONS AND ARTIFICIAL INTELLIGENCE: TECHNOLOGICAL PROMISES AND PRACTICAL REALITIES

VLADISLAV CHERNAVSKIKH*

SUMMARY

● Recent advances in the capabilities of artificial intelligence (AI) have increased state interest in leveraging AI for military purposes. Military integration of advanced AI by nuclear-armed states has the potential to have an impact on elements of their nuclear deterrence architecture such as missile early-warning systems, intelligence, surveillance and reconnaissance (ISR) and nuclear command, control and communications (NC3), as well as related conventional systems.

At the same time, a number of technological and logistical factors can potentially limit or slow the adoption of AI in the nuclear domain. Among these are unreliability of output, susceptibility to cyberattacks, lack of good-quality data, and inadequate hardware and an underdeveloped national industrial and technical base.

Given the current and relatively early stage of military adoption of advanced AI, the exploration of these factors lays the groundwork for further consideration of the likely realities of integration and of potential transparency measures and governance practices at the AI–nuclear nexus.

I. Introduction

The past decade has witnessed a rapid acceleration in the capabilities of artificial intelligence (AI). This process has been driven by advances in machine learning (ML) algorithms, which allow computer systems to 'learn' from data to perform tasks that would otherwise require human intelligence. Increased capabilities in areas such as computer vision, natural language processing, robotics and autonomous systems have, in turn, increased state interest in leveraging AI systems for military purposes. ML-enabled systems, for instance, are being used in ongoing conflicts in Gaza and Ukraine.[1] There are signs that advanced AI may be integrated across the full spectrum of military operations, including for threat monitoring, navigation, precision targeting, intelligence, surveillance and reconnaissance (ISR), decision support, and offensive and defensive cyber operations.[2]

Notably, all nine nuclear-armed states—China, France, Israel, India, the Democratic People's Republic of Korea (DPRK, or North Korea), Pakistan, the Russian Federation, the United Kingdom and the United States—have demonstrated interest in the development and integration of advanced AI capabilities in their militaries, with some explicitly making AI a strategic priority.[3] While traditional AI systems have long been a part of the nuclear weapon enterprise, the potential use of advanced AI in this context could have significant consequences for strategic stability and could increase risks of nuclear conflict. Previous research has identified a variety of ways in which AI could potentially bolster early warning and ISR, nuclear command, control and communications (NC3), delivery systems and conventional systems

[1] Sylvia, N., 'The Israel Defence Forces' use of AI in Gaza: A case of misplaced purpose', Royal United Services Institute (RUSI), 4 July 2024; and Tokariuk, O., 'Ukraine's secret weapon—Artificial intelligence', Center for European Policy Analysis (CEPA), 20 Nov. 2023.

[2] Grand-Clément, S., *Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain* (UNIDIR: Geneva, 2023); and Puscas, I., *AI and International Security: Understanding the Risks and Paving the Path for Confidence-building Measures* (UNIDIR: Geneva, 2023).

[3] Boulanin, V. et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (SIPRI: Stockholm, June 2020); Hoell, M. and Mishra, S., 'Artificial intelligence in nuclear command, control, and communications: Implications for the Nuclear Non-Proliferation Treaty', eds J. Berghofer et al., *The Implications of Emerging Technologies in the Euro-Atlantic Space: Views from the Younger Generation Leaders Network* (Palgrave Macmillan: Cham, 2023); and Borchert, H., Schütz, T. and Verbovszky, J. (eds), *The Very Long Game: 25 Case Studies on the Global State of Defense AI* (Springer: Cham, 2024).

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

with counterforce potential.[4] However, the flurry of recent developments in ML demands reconsideration of the AI–nuclear nexus.

This paper outlines the state of play of the technical possibilities for integration of advanced AI. Section II briefly discusses recent advances in AI capabilities, contextualizing them in technological terms. Section III then considers the degree to which advanced AI is being considered for integration in the nuclear domain. Section IV offers initial analysis of the determinant factors that will drive, or hinder, potential integration. Finally, section V provides some concluding thoughts on the present and future of the AI–nuclear nexus.

II. Situating advances in AI technology

Artificial intelligence is an umbrella term that refers to a wide range of computational methods and approaches that allow machine-based systems to perform intelligent tasks. These tasks can include recognizing patterns, processing natural language, perceiving the environment through computer vision, learning from experience, drawing conclusions and taking action.[5] AI systems are computers or machines that can infer, from the input they receive, how to generate output such as predictions, content, recommendations or decisions in order to complete these tasks.[6] Two primary approaches to creating AI systems are rule-based AI and machine learning, which includes the subfield of deep learning.

Rule-based AI

Rule-based AI has long been integrated in elements of NC3 systems, dating back to its use by the Soviet Union and the USA during the cold war.[7] Predictable and transparent, the operations of these systems are directly tied to developer-set rules. Accordingly, rule-based AI is used for narrow tasks, for instance to optimize sensor data fusion and create robust communications pathways.[8] Missile early-warning systems rely on rule-based AI to identify the launch and trajectory of ballistic missiles with space- or ground-based sensors and transmit this information to human operators for validation.[9] Other applications include automated communication for the secure transmission of orders for missile launches and emergency action messages in the

[4] Geist, E. and Lohn, A. J., *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Rand Corp.: Santa Monica, CA, 2018); Boulanin et al. (note 3); 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).

[5] Boulanin et al. (note 3), pp. 7–8.

[6] While there is currently no universally agreed definition of an AI system, the definition provided here is a simplified version of that in Organisation for Economic Co-operation and Development (OECD), 'Explanatory memorandum on the updated OECD definition of an AI system', OECD Artificial Intelligence Papers no. 8, Mar. 2024, p. 4.

[7] Borrie, J., 'Cold war lessons for automation in nuclear weapon systems', ed. V. Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, vol. I, *Euro-Atlantic Perspectives* (SIPRI: Stockholm, May 2019), pp. 43–46.

[8] McDonnell, T. et al., *Artificial Intelligence in Nuclear Operations: Challenges, Opportunities, and Impacts* (Center for Naval Analyses: Arlington, VA, Apr. 2023), pp. 44–45.

[9] McDonnell et al. (note 8).

event of a nuclear attack, as well as assistance with rapid missile targeting and missile guidance.[10]

Machine learning and deep learning

While rule-based AI systems are known to function consistently, without interruption and quickly (especially in comparison to human beings), their rules require continuous and extensive updates in order to adapt to new situations. This limits their applicability.[11] In contrast, ML focuses on creating AI systems that are capable of 'learning' how to produce outputs from new data without explicit programming.[12] ML systems learn through a training process in which mathematical algorithms are iteratively applied to extensive data sets to identify statistical patterns and associate them with expected outputs.[13] The learned statistical patterns, known as parameters, form the core of an ML model—a type of program that acts as the 'brain' of an AI system.[14] The model can then apply these parameters to new, previously unseen data to generate output in a process called inference. In some cases, the model continues to adjust the parameters even after it has been deployed in the operating environment.[15]

In the past decade, an ML approach called deep learning (DL) has become the most popular way to train AI models. It uses complex algorithms called deep neural network architectures that are capable of processing larger amounts of data faster and can capture more complex statistical relationships.[16] DL serves as the backbone of the recent rapid advance in AI capabilities.[17] It has enhanced the performance of AI systems in a number of areas.[18] These include tasks that require recognizing complex patterns, such as in computer vision (e.g. recognizing and classifying objects, people or scenes), natural language processing (e.g. voice and speech recognition) and signal recognition (e.g. acoustic or electromagnetic signatures). It has also included data-management tasks, such as the fusion of data from different sources, its organization and analysis.[19]

[10] Horowitz, M. C., Scharre, P. and Velez-Green, A., 'A stable nuclear future? The impact of autonomous systems and artificial intelligence', arXiv 1912.05291, 13 Dec. 2019; and Baldus, J., 'Doomsday machines? Nukes, nuclear verification and artificial intelligence', eds T. Reinhold and N. Schörnig, *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm* (Springer: Cham, 2022).

[11] Hruby, J. and Miller, M. N., 'Assessing and managing the benefits and risks of artificial intelligence in nuclear-weapon systems', Nuclear Threat Initiative, Aug. 2021; and DataCamp, 'What is symbolic AI?', Artificial Intelligence Blog, May 2023.

[12] Hurwitz, J. and Kirsch, D., *Machine Learning for Dummies*, IBM Limited Edition (John Wiley and Sons: Hoboken, NJ, 2018).

[13] Eddine Abail, I. et al., 'Artificial intelligence & machine learning', Technology Primers for Policymakers, Harvard University, Belfer Center for Science and International Affairs, Apr. 2023.

[14] Hurwitz and Kirsch (note 12).

[15] International Committee of the Red Cross (ICRC), 'What you need to know about artificial intelligence in armed conflict', 6 Oct. 2023.

[16] Amazon Web Services, 'What is a neural network?', [n.d.].

[17] Case Western Reserve University, 'Advancements in artificial intelligence and machine learning', Online Engineering Blog, 25 Mar. 2024.

[18] Holdsworth, J. and Scapicchio, M., 'What is deep learning?', IBM, 17 June 2024.

[19] Boulanin et al. (note 3).

Foundation models

In 2017 Google introduced a new deep neural network architecture called a transformer that could process larger amounts of data more quickly and could capture more complex relationships within it. Transformers are scalable, which means that the precision, efficiency and complexity of the output of the model predictably improve with increases in the amount of training data and computational resources.[20] DL and transformer architectures have thus enabled the development of flexible, general-purpose AI models that can effectively perform a wider range of complex tasks.[21] These large-scale models, also known as foundation models or general-purpose AI, represent a significant leap forwards.[22]

Foundation models are trained on massive amounts of data in a process called pre-training to capture a broad range of statistical patterns and relationships. A pre-trained model can be expediently fine-tuned on a smaller data set to enable it to specialize in a specific domain.[23] Fine-tuning requires significantly less data than pre-training, which means that particular capabilities of these models can change at relatively low cost. Nonetheless, the training and operation of foundation models require significant computational resources—hardware (e.g. graphic processing units, GPUs), software platforms (e.g. cloud computing) and infrastructure (e.g. data centres).[24] Data sets used for training foundation models are critically important as they directly influence the accuracy, effectiveness and reliability of the model after deployment.

Foundation models include large language models (LLMs) such as GPT (for Generative Pre-trained Transformer), which powers ChatGPT, and other types of generative AI that can produce original content in the form of text, images, audio and video. In computer vision, foundation models enable more sophisticated image classification, object detection, image generation and video analysis.[25] The innovative nature of transformer architecture has further enabled development of multimodal AI, which can process and analyse different types of input data simultaneously and produce outputs that may differ from the input source type.[26]

III. Integration of advanced AI in the nuclear domain

Recent improvements in DL have spurred among nuclear-armed states and their allies a new wave of interest and government investment in exploring options for advanced applications of AI in the military sector.[27] Integration in a wide range of capabilities is currently technically feasible across the military

[20] Kaplan, J. et al., 'Scaling laws for neural language models', arXiv 2001.08361, 23 Jan. 2020.

[21] Harris, J., Harris, E. and Beall, M., 'Survey of AI technologies and AI R&D trajectories', Gladstone AI, 3 Nov. 2023.

[22] Madiega, T. A., 'General-purpose artificial intelligence', At a Glance, European Parliamentary Research Service, Mar. 2023.

[23] Amazon Web Services, 'What are foundation models?', [n.d.]; and Hickey, A., 'The GPT dilemma: Foundation models and the shadow of dual-use', arXiv 2407.20442, 29 July 2024.

[24] Gupta, A. and Ranjan, A., 'A primer on compute', Carnegie India, 30 Apr. 2024.

[25] Awais, M. et al., 'Foundational models defining a new era in vision: A survey and outlook', arXiv 2307.13721, 25 July 2023.

[26] Amazon Web Services, 'What are transformers in artificial intelligence?', [n.d.].

[27] eds Borchert et al. (note 3).

domain—from weapon-specific applications (e.g. improved targeting and missile guidance) to non-weapon-related uses (e.g. fast data collection, fusion and analysis, as well as simulations).[28] Given the culture of secrecy surrounding nuclear weapon programmes, information on plans, strategies or practices related to further AI–nuclear integration is severely limited. Nonetheless, the discussion of AI in the broader defence domain can provide some clues as to where states see the value of these systems, including for the nuclear domain. It also suggests implications for nuclear deterrence.

In the nuclear context, ML is often discussed as having the potential to enhance capabilities in all elements of the nuclear deterrence architecture: early warning and ISR; command and control; nuclear weapon targeting, guidance and navigation; and non-nuclear operations such as missile defence, cybersecurity and anti-submarine warfare.[29] For instance, India is planning to launch a fleet of 50 AI-enabled surveillance satellites and is investigating the use of DL in radar technology.[30] Meanwhile, the US National Geospatial-Intelligence Agency plans significant investment in data-labelling services over five years to enhance ML capabilities for analysing satellite imagery and geospatial data.[31] Further ML-enabled improvements in ISR capabilities might, for instance, aid in detection, tracking and targeting of nuclear delivery systems, whether missile silos, aircraft or mobile launchers.[32]

Beyond this, the US Department of Defense (DOD) is examining how to leverage advanced AI to enable its missile defence systems to effectively track and engage cruise and hypersonic missiles as well as analyse data collected by low earth orbit sensors.[33] It has also reportedly invested in using AI to manage large data sets, optimize testing, and improve threat detection and engagement strategies broadly across the US Missile Defense Agency (MDA).[34] When it comes to Russia, available information is limited, but studies also suggest potential interest in the use of ML and DL in air and missile defence forces.[35] For example, an AI control capability is reportedly being developed for the S-500 air-defence system, which is designed to counter aircraft and ballistic and cruise missiles and which can supposedly target low earth orbit satellites.[36]

[28] Grand-Clément (note 2), pp. 14–15; and Unal, B. and Richard, U., *Governance of Artificial Intelligence in the Military Domain*, UNODA Occasional Papers no. 42 (UNODA: New York, June 2024), pp. 13–14.

[29] McDonnell et al. (note 8); Boulanin et al. (note 3); and Hruby and Miller (note 11).

[30] Singh, S., 'ISRO plans 50 AI-based surveillance satellites', *Times of India*, 30 Dec. 2023; Thakur, S. P., 'Integrating AI in India's defence sector', Bloomsbury Intelligence & Security Institute, 12 June 2024; and Indian Ministry of Defence, Department of Defence Production (DDP), *Artificial Intelligence in Defence: The New Age of Defence* (DDP: New Delhi, July 2022).

[31] Erwin, S., 'NGA to launch $700 million program to help AI make sense of satellite images', *SpaceNews*, 3 Sep. 2024.

[32] van Hooft, P., 'AI and nuclear weapons: Keeping the human in the loop, not only for the decision, but also before the decision', The Hague Center for Strategic Studies, 3 Mar. 2024.

[33] Freedberg, S. J., 'MDA launches missile defense battle management upgrade with $847M order to Lockheed Martin', Breaking Defense, 12 Apr. 2024; and Erwin, S., 'AI company developing software to detect hypersonic missiles from space', *SpaceNews*, 18 Feb. 2024.

[34] 'Enterprise AI applications ordered under US MDA production agreement', Defense Advancement, 19 Dec. 2022.

[35] McDonnell et al. (note 8).

[36] Starchak, M., 'Russian defense plan kicks off separate AI development push', *Defense News*, 16 Aug. 2024.

The naval domain appears to be a particular area of focus for the development and deployment of ML systems. The underwater use of ML and DL is significant as a more transparent sea can threaten the survivability of nuclear submarines, potentially upending strategic stability.[37] In 2024 the US Navy and the Defense Innovation Unit announced an integrated capability for underwater threat detection that was developed in just one year with the use of commercial ML technology integrated into existing uncrewed underwater vehicles (UUVs).[38] Australia, the UK and the USA, through their AUKUS agreement, are investigating how AI can track Chinese submarines with greater speed and accuracy by accelerating the processing of underwater acoustic signals and sonar data.[39] France is also doing this.[40] For its part, the People's Liberation Army (PLA) Navy—China's navy—has invested in a UUV designed specifically for hunting submarines as part of a greater self-declared emphasis on 'mechanization, informatization and intelligentization'.[41] The release of a data set of Chinese origin has also led some experts to conclude that the PLA is probably already also using DL systems for ship detection.[42]

The above examples suggest that many military AI capabilities currently being deployed or planned for deployment emphasize adaptation and reaction to changing information for tasks such as data fusion, image classification and analysis.[43] However, applications of advanced AI for nuclear weapon delivery are also being explored. As of February 2024 an autonomous nuclear-armed UUV called Poseidon that is being developed by Russia was 'about to complete its testing stage' according to Russian President Vladimir Putin.[44] North Korean leader Kim Jong Un has similarly urged incorporation of AI in nuclear torpedoes and other UUVs under development.[45] Meanwhile, the USA has started the production of the B-21 Raider, a nuclear-capable long-range bomber that is designed to undertake crewed or uncrewed operations in collaboration with other crewed and uncrewed aircraft.[46] The Pakistan Air Force, which operates nuclear-capable aircraft, has its own dedicated research centre that is reportedly exploring applications of ML and DL, predictive analysis, and natural language processing, although no reliable information is available on whether these would directly concern nuclear-delivery platforms.[47]

[37] Erästö, T., Su, F and Wan, W., *Navigating Security Dilemmas in Indo-Pacific Waters: Undersea Capabilities and Armament Dynamics* (SIPRI: Stockholm, June 2024).

[38] US Defense Innovation Unit, 'DOD successfully deploys commercial AI infrastructure to support underwater target threat detection', 17 June 2024.

[39] Freedberg, S. J., 'Transparent sea: AUKUS looks to AI, quantum in hunt for Chinese submarines', Breaking Defense, 29 Jan. 2024.

[40] Ruitenberg, R., 'France turns to AI for signals analysis in underwater acoustics war', C4ISRNET, 17 May 2024.

[41] Liu, X., 'China displays land, sea, air combat robots at expo', *Global Times*, 5 July 2021.

[42] Gupta, R. et al., Berkeley Risk and Security Lab, 'Open-source assessments of AI capabilities: The proliferation of AI analysis tools, replicating competitor models, and the Zhousidun dataset', arXiv 2405.12167, 24 May 2024.

[43] Grand-Clément (note 2).

[44] Putin, V., Presidential address to the Federal Assembly, 29 Feb. 2024.

[45] Zwirko, C., 'Kim Jong Un inspects new "suicide drones", urges incorporation of AI', NK News, 26 Aug. 2024.

[46] Losey, S., 'Pentagon OKs B-21 for low-rate production after successful tests', *Defense News*, 23 Jan. 2024; and Lopez, C. T., 'World gets first look at B-21 Raider', US Department of Defense, 3 Dec. 2022.

[47] Ali, U., 'Comparing the AI–military integration by India and Pakistan', Centre for Strategic and Contemporary Research, 7 Sep. 2023.

Furthermore, there are indications that nuclear-armed states are already exploring ways in which foundation models and generative AI may be used in military systems. For instance, in 2023 the US DOD launched a dedicated project, Task Force Lima, to investigate the possibilities of integrating generative AI and LLMs in the military sector. Branches of the US military are experimenting with custom LLMs to perform routine data-analysis tasks such as summarizing and classifying text, as well as for wargaming.[48] China reportedly employs Baidu's Ernie Bot, an LLM similar to ChatGPT, to enhance combat simulations and support decision-making.[49] It is also reported to be exploring possible applications of generative AI for cyber-enabled influence operations.[50] The British Ministry of Defence (MOD) is cooperating with the private sector to develop training simulations that are enhanced by generative AI and to investigate capabilities of LLMs in cyber defence.[51] France's Alternative Energies and Atomic Energy Commission (Commissariat à l'énergie atomique et aux énergies alternatives, CEA), a government-funded research organization, is working with Thales, a private French electronics and armaments manufacturer, to develop LLMs and vision–language generative models specifically for military intelligence gathering and the acceleration of command and control processes through data analysis, as well as for increasing interoperability between military allies.[52]

It would be inaccurate to suggest that any of the above constitutes a fully operational advanced AI capability, given their ongoing development and the fairly narrow tasks that newer ML systems are being assigned. Yet the fact that some capabilities are already being experimented with suggests clear trends in the direction of further integration. For instance, according to some assessments, there is a consensus in the PLA media that generative AI has a place in warfare, including in human–machine interaction, decision-making, network warfare, logistics, the cognitive domain, the space domain and training.[53] Any of these uses is likely to have an impact on the environment in which nuclear weapons operate. This is true even with technologies that remain exclusively in the private sector; for instance, Rhombus Power, a US company that describes itself as an 'AI digital nervous system for defense and national security', claims that it has alerted unnamed US government customers to imminent missile launches by North Korea and space operations by China.[54] Another example is the French company Preligens, which focuses on AI applications for detection and identification of military assets

[48] Caballero W. N. and Jenkins P. R., 'On large language models in national security applications', arXiv 2407.03453, 3 July 2024.

[49] Caballero and Jenkins (note 48), p. 6.

[50] Hickey (note 23); and Beauchamp-Mustafaga, N., *Exploring the Implications of Generative AI for Chinese Military Cyber-enabled Influence Operations: Chinese Military Strategies, Capabilities, and Intent*, Testimony presented before the US–China Economic and Security Review Commission (Rand Corp.: Santa Monica, CA, 1 Feb. 2024).

[51] Simpson, S., 'British Army training simulations to be enhanced by generative AI', Defense Advancement, 5 Feb. 2024; and British Ministry of Defence, Defence Science and Technology Laboratory (DSTL), 'DSTL and Google Cloud hackathon: New era of defence AI innovation', 20 Nov. 2023.

[52] Thales Group, 'Thales and CEA partner on trusted generative AI for defence and security', 17 June 2024.

[53] Baughman, J., 'China's ChatGPT war', US Department of the Air Force, China Aerospace Studies Institute, 21 Aug. 2023.

[54] Bajak, F., 'US intelligence agencies' embrace of generative AI is at once wary and urgent', PBS, 24 May 2024.

with commercial and government satellite imagery. Preligens has a contract for data-processing software with the French MOD's Directorate General of Armaments, supplies software to the North Atlantic Treaty Organization (NATO) and provides AI for optical sensor analysis in the USA.[55]

Notwithstanding the above developments, integration of advanced AI into the military domain should not be seen as inevitable, and it will certainly not be ubiquitous.

IV. Challenges in integrating advanced AI in the nuclear domain

A host of factors may limit or slow the adoption of advanced AI capabilities, including foundation models, in the military domain broadly and in the nuclear domain in particular. Some of these are technical, arising from inherent challenges in the technology, and are not limited to nuclear integration. Other factors pertain specifically to the act of integration, while a third set relates to logistical difficulties that a state can face even if it wants to pursue integration. The latter centre on access to the resources necessary for the development and deployment of advanced AI.

Technical challenges

Foundation models suffer from several technical drawbacks. These are also to some extent present in broader ML and DL approaches.

To begin with, advanced AI models suffer from unreliability—that is, a lack of trustworthiness. They have consistently been shown to hallucinate, meaning that they can confidently produce outputs that are incorrect and are unsupported by their training data.[56] This can mean that an LLM model invents facts or that a large-scale vision model incorrectly identifies an object in an image, leading to inaccurate assessments or false positives in critical areas such as threat detection and surveillance. This has led the chief technology officer of the US Central Intelligence Agency (CIA) to suggest treating generative AI as a 'crazy, drunk friend'.[57] More advanced AI capabilities have enabled improved analysis, predictions of behaviour or evolutions of certain scenarios. However, performance has come at the expense of interpretability. The more parameters a ML model has, the harder it is to trace how particular inputs lead to specific outputs. Foundation models, for instance, typically have billions of parameters.[58] The scale and complexity make it hard to interpret or understand the specific reasoning behind their output, creating a black box problem. The lack of transparency (and accompanying difficulties in troubleshooting) complicates efforts to verify AI-generated predictions in critical decision-making scenarios, in which policymakers are likely to be

[55] Werner, D., 'Preligens aims to become a long-term DOD supplier', *SpaceNews*, 24 Oct. 2022.

[56] Saltini, A., European Leadership Network, 'The risks of AI integration with NC3', Written evidence (AIW0023), British Parliament, House of Lords, AI in Weapon Systems Committee, 8 June 2023.

[57] Bajak (note 54).

[58] Amazon Web Services (note 23).

under significant time pressure.[59] Distrust of the output produced by AI can lead human operators to, for example, disregard correctly aggregated ISR data; conversely, over-reliance on AI output can, for example, lead a human operator to make a decision to escalate without additional verification in a scenario where ML is used to enhance missile early-warning systems.[60]

Lack of trustworthiness and interpretability of foundation models can contribute to further misalignment between an operator's intent and a system's actual behaviour; to a system's lack of resilience to unusual situations or events (robustness); to an operator's inability to detect unexpected model outcomes or functionalities (monitoring); or to problems that emerge from the broader context in which AI systems are handled (systemic safety).[61] While techniques exist to make AI more explainable, this has so far resulted in a trade-off in performance.[62] Further technological developments may resolve problems of interpretability and transparency but the resource problem (explored below) looms.[63]

Further, advanced AI systems are particularly susceptible to cybersecurity threats, which provide adversaries and non-state actors with opportunities to compromise them.[64] ML and DL models can be deceived into producing faulty output through adversarial manipulation of training data (data poisoning) or input data fed to an AI system in real-time (input manipulation). This can lead, for example, to identification mistakes in ISR tasks. Hackers can also attempt to extract information about an AI system's operations or training data, thereby exposing the classified data on which military AI models rely.[65] While ML-enabled intrusion detection can offset some of these effects, defensive measures against such cyberthreats are generally inadequate.[66] Furthermore, strengthening defences against adversarial attacks often compromises the ability of an ML system to detect novel threats, which creates a critical trade-off when trying to secure such a system.[67] Such risks are especially concerning given the widespread development and increased sophistication of offensive cyber capabilities and operations.[68]

The technical challenges facing AI are well recognized among nuclear-armed states. For instance, ensuring safety and reliability of AI capabilities is one of the main principles underpinning stated policy approaches to 'respon-

[59] Grand-Clément (note 2); Saltini (note 4); and Zala, B., 'Should AI stay or should AI go? First strike incentives & deterrence stability', *Australian Journal of International Affairs*, vol. 78, no. 2 (2024), pp. 157–58.

[60] McDonnell et al. (note 8).

[61] Hoffman, W. and Kim, H. M., *Reducing the Risks of Artificial Intelligence for Military Decision Advantage* (Georgetown University, Center for Security and Emerging Technology: Washington, DC, Mar. 2023).

[62] Ali, S. et al., 'Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence', *Information Fusion*, vol. 99 (Nov. 2023).

[63] Templeton, A. et al., 'Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet', Transformer Circuits Thread, 21 May 2024.

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

[65] Saltini (note 64); and the British Department for Science, Innovation and Technology, 'Cyber security risks to artificial intelligence', 15 May 2024.

[66] Saltini (note 64); and Vassilev, A. et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* (National Institute of Standards and Technology: Gaithersburg, MD, Jan. 2024).

[67] Hoffman, W., *Making AI Work for Cyber Defense: The Accuracy–Robustness Tradeoff* (Georgetown University, Center for Security and Emerging Technology: Washington, DC, Dec. 2021).

[68] Hruby and Miller (note 11), pp. 13–14.

sible' adoption of AI in the military domain by the UK and the USA.[69] In this regard, Kathleen Hicks, the US deputy defence secretary, stated that most commercially available systems enabled by LLMs are not yet technically mature enough to comply with DOD principles.[70] Russian MOD researchers argue that the lack of transparency, interpretability, robustness and controllability of current AI models presents a major constraint on integration of AI in Russia's Strategic Rocket Forces.[71] Chinese defence experts similarly posit that the PLA's current inability to develop sufficiently trustworthy AI systems presents a roadblock to military AI in China.[72]

Integration challenges

The technical challenges discussed above are not exclusive to integration in the nuclear domain. However, they will feature prominently in the risk calculus of the policymakers who are considering further AI–nuclear integration. They are especially prominent given the central importance of nuclear weapon programmes in national security and the enormous consequences—strategic, operational, political and humanitarian, among others—of nuclear weapon use. Some experts argue that the consequences of failure mean that states are likely to continue relying on rule-based AI rather than shifting to advanced AI capabilities in the nuclear domain.[73] Indeed, for military uses, attributes like durability, security and reliability are often more crucial than speed. Existing AI models used by militaries tend to employ traditional ML with significantly lower computational requirements, tailored to the specific demands of military environments. Even as militaries adopt foundation models, these systems are unlikely to use leading-edge microchips.[74] Integrating foundation models with existing military legacy systems presents a significant challenge, as it is likely to require extensive modifications to those systems.[75]

Lack of good-quality data can hamper the efficiency and accuracy of foundation models. This includes limited training data sets, which can have an impact on the ability of AI algorithms to operate in real-world settings; incoherence or fragmented data because there are too few data-capture methods (e.g. via sensors); and poor data-management systems whereby data-sharing is hampered or there is poor data hygiene.[76] Problems may also emerge if data collection methods are not able to keep up with fast-paced changes happening on the ground. In cases such as these, new information

[69] British Ministry of Defence (MOD), *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-enabled Capability in Defence* (MOD: London, June 2022); and US Department of Defense (DOD), *US Department of Defense Responsible Artificial Intelligence Strategy and Implementation Pathway* (DOD: Washington, DC, June 2022).

[70] Hicks, K., US Deputy Secretary of Defense, Press briefing, 2 Nov. 2023.

[71] Shakirov, O., *Russian Thinking on AI Integration and Interaction with Nuclear Command and Control, Force Structure, and Decision-making* (European Leadership Network: London, Nov. 2023), pp. 5–6.

[72] Bresnick, S., *China's Military AI Roadblocks: PRC Perspectives on Technological Challenges to Intelligentized Warfare* (Georgetown University, Center for Security and Emerging Technology: Washington, DC, June 2024), pp. 17–18.

[73] Hruby and Miller (note 11), p. 6.

[74] Hickey (note 23), p. 10.

[75] Chochtoulas, A., 'How large language models are transforming modern warfare', *Transforming Joint Air & Space Power: Journal of the Joint Air Power Competence Centre*, no. 37 (May 2024).

[76] Grand-Clément (note 2), pp. 27–28.

can be missed or delayed, resulting in incorrect assessments or incomplete situational understanding. Finally, the use of AI to help with scenarios and simulations such as wargaming could run the risk of mirroring, whereby the AI system is only capable of repeating known information but cannot provide new insights.[77]

All of these limitations and concerns feature prominently in the nuclear domain.[78] The relative lack of data points relating to real-life examples of nuclear crises and the use of nuclear weapons (especially those that reflect the contemporary multipolar nuclear landscape) creates a greater risk of errors and unpredictable results in situations where advanced AI may be applied for decision support in command and control.[79] In the same sense, the amount of information available on the composition of nuclear forces and on NC3 systems and practices differs significantly between nuclear-armed states. While some suggest that new AI foundation models can be extended with a corpus of classified nuclear weapon data, limited access to that data even within national security circles can exacerbate the black box problem and also raises the question of accountability for decision-making.[80] Finally, the practice of strategic ambiguity in declaratory nuclear policy—stating intentions and red lines clearly enough to deter attacks, but not so explicitly as to restrict freedom of action or encourage adversary aggression—adds additional uncertainty with which AI decision-support systems might struggle.[81]

The expectations placed on AI are high for certain tasks—and perhaps unrealistic, at least in the near-term future. While AI does, in some regards, offer capabilities that exceed those of humans, there are still barriers. Notable among these barriers are limitations of AI itself and limitations in the amount of data available, both for training but also in practice. There is an expectation that AI will be more accurate in its analysis of data than human-led efforts in the area or will be able to develop superior tactics. However, this is a question of access: to data, systems and equipment.

Resource access challenges

The notion of sovereign AI—a state's capability to indigenously develop its own foundation models—has received increased attention in recent years.[82] The value of secure AI models and infrastructure becomes critical in the nuclear domain due to the high stakes involved. Any state pursuing integration of advanced AI into its nuclear deterrence architecture will need to ensure the highest degree of reliability and accuracy of the output of the AI model while safeguarding its AI systems against cyberattacks or other subversions (e.g. training data poisoning or input manipulation).[83] Competition

[77] Grand-Clément (note 2), pp. 27–28.

[78] McDonnell et al. (note 8), pp. 18–21.

[79] Macartney, S., 'With AI, regulations must come before benefits', Nukes of Hazard, Center for Arms Control and Non-Proliferation, 17 Aug. 2023.

[80] US National Nuclear Security Administration (NNSA), *Artificial Intelligence for Nuclear Deterrence Strategy 2023* (NNSA: Washington, DC, 2023).

[81] Johnson (note 4), pp. 184–87.

[82] Chavez, P., 'The rising tide of sovereign AI', Center for a New American Security, 10 Feb. 2024; and Hajkowicz, S. A., *Artificial Intelligence Foundation Models: Industry Enablement, Productivity Growth, Policy Levers and Sovereign Capability Considerations for Australia* (CSIRO: Canberra, Mar. 2024).

[83] Geist and Lohn (note 4), pp. 19–20.

will centre on acquisition of the resources necessary to develop and integrate advanced AI capabilities. These include the key performance drivers of AI: talent, large volumes of high-quality data to train ML systems and powerful computational resources.[84] Yet, as discussed above, data might not be readily defined or available, especially in nuclear-armed states with smaller arsenals or less mature nuclear weapon programmes. Further, the development of AI capabilities is driven by a small number of companies located largely in the USA.[85] The prominence of foundation models in the industry sector underscores the capacity of such US companies as Microsoft, Meta, OpenAI, Anthropic and Google to bear the computational costs of training on vast volumes of data.[86] As of the end of 2023, most foundation models were created in the USA (109), followed by China (20) and the UK (8).[87] Most of these models come from industry players and these three nuclear-armed states plan to rely on strategic collaboration between the civilian and military sectors to advance AI technologies and applications.[88]

A significant limitation to resource access is the concentration of the specialized computing hardware necessary to handle the computational demands of building foundation models. AI capabilities rely largely on the processing power of specialized microchips such as modern GPUs, tensor processing units (TPUs) and neural processing units (NPUs), which are expensive and difficult to produce.[89] Nearly all—92 per cent—of the total value of the global semiconductor supply chain is located in the USA (at 39 per cent) and allied states and regions—Western Europe (especially Germany, the Netherlands and the UK), Japan, the Republic of Korea (South Korea) and Taiwan (together contributing 53 per cent).[90] Furthermore, Japan and the Netherlands are the only providers of the type of photolithography equipment necessary for producing advanced AI chips.[91] Access to advanced chips represents an obvious bottleneck for nuclear-armed states such as Russia, which is unable to produce sophisticated microelectronic components domestically and relies on imports to support its AI development.[92]

[84] Boulanin et al. (note 3), p. 32; and Buchanan, B., *The AI Triad and What It Means for National Security Strategy* (Georgetown University, Center for Security and Emerging Technology: Washington, DC, Aug. 2020).

[85] Kak, A. and Myers West, S., *AI Now 2023 Landscape: Confronting Tech Power* (AI Now Institute: New York, 11 Apr. 2023).

[86] Maslej, N. (ed.), AI Index Steering Committee, *Artificial Intelligence Index Report 2024* (Stanford University, Institute for Human-Centered AI: Stanford, CA, Apr. 2024).

[87] ed. Maslej (note 86), p. 61.

[88] ed. Maslej (note 86), p. 46; Dahlgren, M., 'Defense priorities in the open-source AI debate: A preliminary assessment', Georgetown University, Center for Strategic & International Studies (CSIS), Aug. 2024; Licata, N. R., 'China's military–civil fusion strategy: A blueprint for technological superiority', Foreign Policy Research Institute, 19 Dec. 2023; and British Government, 'Defence Artificial Intelligence Centre', [n.d.].

[89] Vipra, J. and Myers West, S., *Computational Power and AI* (AI Now Institute: New York, 27 Sep. 2023); and Sevilla, J. et al., 'Compute trends across three eras of machine learning', arXiv 2202.05924, 9 Mar. 2022.

[90] Khan, S. M., Peterson, D. and Mann, A., *The Semiconductor Supply Chain: Assessing National Competitiveness* (Georgetown University, Center for Security and Emerging Technology: Washington, DC, Jan. 2021).

[91] Khan et al. (note 90), p. 30.

[92] Kossek M. and Stewart I., 'How (and how many) Western chips are getting to Russia?', TradeCompliance.io, James Martin Center for Nonproliferation Studies, 26 Aug. 2024; and Byrne, J. et al., *Silicon Lifeline: Western Electronics at the Heart of Russia's War Machine* (Royal United Services Institute: London, Aug. 2022).

Poor geopolitical relations may further limit access. For instance, the USA has imposed strict controls on exports of advanced semiconductor-manufacturing technology to China, and there were reports in 2023 that it agreed with Japan and the Netherlands to further restrict China's access to chipmaking tools.[93] In an effort to lessen its dependence on these states, China aims to strengthen domestic chipmaking capabilities and expand its national computing power network.[94] Likewise, the Western sanctions imposed on Russia after February 2022 disrupted its access to high-quality foreign hardware and components; this resulted in the Russian government scaling back the scope of its national AI-development road map and pivoting to the goal of first building an indigenous production base for microelectronics, including specialized AI chips.[95]

Building a secure and effective system of AI data centres to process training data requires additional infrastructure for cloud computing and low-latency data transfer. Given all of the above, advanced AI integration simply may not be economically viable for some nuclear-armed states or may not contain sufficient operational value to justify pursuing in the nuclear domain.

V. Conclusions

There is an overall lack of reliable information on the particular types of AI technology that nuclear-armed states are pursuing. However, it is undeniable that advanced AI based on deep learning is seen by most of these states as a technology with the potential to be integrated across the military spectrum. Many of the applications currently being explored by nuclear-armed states focus on data fusion, analysis and simulations in such areas as ISR (for enhanced targeting and anti-submarine warfare) and missile defence. The role of generative AI is also being explored by nuclear-armed states for decision-support tasks and military planning, as well as defensive and offensive cyber operations.

However, a number of factors have the potential to limit the prospects of AI–nuclear integration. DL and foundation models are still plagued by inherent technical problems. Unreliability in their output as well as susceptibility to cyberattacks compounded by the black box nature of these systems can lead to critical failures when integrated in sensitive nuclear weapons and adjacent systems. Lack of good-quality data in the nuclear domain can further hamper the efficiency and accuracy of advanced AI models. Finally, development of advanced AI requires access to specialized computing hardware, high-quality data, and an adequate industrial and technical base.

[93] E.g. Zhou, J., Su, F. and Yuan, J., 'De-risking: The EU's and Japan's approaches to managing economic relations with China', SIPRI, Feb. 2024.

[94] 'China aims to increase computing power by more than 50% by 2025', *Global Times*, 9 Oct. 2023; Arcesati, R., 'China's AI development model in an era of technological deglobalization', Mercator Institute for China Studies (MERICS) and UC Institute on Global Conflict and Cooperation (IGCC), 2 May 2024; and Wang, C., '4 ways China gets around US AI chip restrictions', The Diplomat, 28 June 2024.

[95] Bendett, S., *The Role of AI in Russia's Confrontation with the West* (Center for a New American Security: Washington, DC, Apr. 2024).

Abbreviations

AI	Artificial intelligence
DL	Deep learning
DOD	Department of Defense (United States)
GPT	Generative Pre-trained Transformer
GPU	Graphics processing unit
ISR	Intelligence, surveillance and reconnaissance
LLM	Large language model
ML	Machine learning
MOD	Ministry of Defence
NC3	Nuclear command, control and communications
PLA	People's Liberation Army (China)
UUV	Uncrewed underwater vehicle

RECENT SIPRI PUBLICATIONS

Climate Change Adaptation in Areas Beyond Government Control: Opportunities and Limitations

Dr Karen Meijer and Ann-Sophie Böhle
September, 2024

From Conflict to Collaboration: Co-funding Environmental Peacebuilding in South-central Somalia

Kheira Tarif
September 2024

Strengthening Social Cohesion in the Nineveh Plains of Iraq: Issues of Common Concern and Local Cooperative Solutions

Amal Bourhrous, Emelie Poignant Khafagi and Dr Alaa Tartir
August 2024

Towards a Two-tiered Approach to Regulation of Autonomous Weapon Systems: Identifying Pathways and Possible Elements

Laura Bruun
August 2024

Cyber Risk Reduction in China, Russia, the United States and the European Union

Dr Lora Saalman, Fei Su and Larisa Saveleva Dovgal
June 2024

Navigating Security Dilemmas in Indo-Pacific Waters: Undersea Capabilities and Armament Dynamics

Dr Tytti Erästö, Fei Su and Dr Wilfred Wan
June 2024

Reducing the Role of Nuclear Weapons in Military Alliances

Dr Tytti Erästö
June 2024

SIPRI is an independent international institute dedicated to research into conflict, armaments, arms control and disarmament. Established in 1966, SIPRI provides data, analysis and recommendations, based on open sources, to policymakers, researchers, media and the interested public.

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

SIPRI BACKGROUND PAPER

NUCLEAR WEAPONS AND ARTIFICIAL INTELLIGENCE: TECHNOLOGICAL PROMISES AND PRACTICAL REALITIES

VLADISLAV CHERNAVSKIKH

CONTENTS

ABOUT THE AUTHOR

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 working group on nuclear disarmament in policy and international law of the Alva Myrdal Centre for Nuclear Disarmament (AMC) at Uppsala University. Prior to joining SIPRI, Chernavskikh worked at the Center for Energy and Security Studies (CENESS), Moscow.