EUROPEAN LEADERSHIP NETWORK

Assessing the implications of integrating AI in nuclear decision-making systems

Policy brief

Alice Saltini

February 2025

This document has been produced with the financial assistance of the EU. The contents are the sole responsibility of the EU Non-Proliferation and Disarmament Consortium promoting the European Network of Independent Non-proliferation and Disarmament Think Tanks, and can in no way be taken to reflect the views of the EU.

This paper was originally prepared as a policy brief for an ad-hoc EU Non-Proliferation and Disarmament Consortium seminar and is being republished with the permission of the European External Action Service (EEAS).

About the Author

Alice Saltini
Former Policy Fellow, European Leadership Network (ELN) and Non-Resident Expert on AI at the James Martin Center for Nonproliferation Studies (CNS)

Alice Saltini is a Non-Resident Expert on AI at the James Martin Center for Nonproliferation Studies (CNS), specialising in the impact of AI on nuclear decision-making. She advises governments and international organisations on managing AI-related nuclear risks, focusing on mitigating the challenges of integrating AI into military and nuclear weapons systems by translating complex technical concepts into actionable policy insights. She has published extensively on military applications of AI and has developed a general-purpose risk assessment framework for analysing AI and nuclear risks.

Previously, she worked with the European Leadership Network, the Comprehensive Nuclear-Test-Ban Treaty Organization, and CNS. She holds a Master's degree in Russian Studies and a Postgraduate Certificate (PgCert) in Nonproliferation Studies from the Middlebury Institute of International Studies.

Summary

This policy brief analyses the integration of artificial intelligence (AI) into nuclear command, control and communications systems(NC3), exploring potential benefits and significant risks. While cautious AI integration can have some benefits for enhancing intelligence collection and situational awareness by automating processes and analysing vast amounts of data, it presents grave risks due to its unreliability, opacity, susceptibility to cyber threats and potential misalignment with human values. Many of the risks and benefits are heavily interconnected as technological attributes directly affect how AI functions in nuclear operations, particularly in decision-making processes. This, in turn, affects states' perceptions as well as the countermeasures they might employ, and ultimately, the balance of these elements determines how deterrence calculations shift.

This paper highlights the need for a better assessment of risks and the establishment of thresholds for integration to prevent miscalculations and nuclear escalation, leading to potentially catastrophic outcomes. It proposes that the European Union leads international dialogue on AI risks in the nuclear domain in relevant international discussions, particularly at the REAIM Summits, integrates AI discussions into the Non-Proliferation Treaty framework, and commissions research to identify and manage high-risk AI applications.

It recommends that the European Union:

- **Leads international dialogue on AI risks in the nuclear domain in relevant international discussions, particularly at the REAIM Summits.**
- **Integrates AI discussions into the Non-Proliferation Treaty framework.**
- **Commissions research to identify and manage high-risk AI applications.**

Many of the risks and benefits are heavily interconnected as technological attributes directly affect how AI functions in nuclear operations, particularly in decision-making processes.

1. Introduction

The current debate on artificial intelligence (AI) and its implications for the military domain has garnered considerable worldwide attention. The issues surrounding lethal autonomous weapon systems (LAWS) have dominated the discussions so far, but the implications of AI in the field of nuclear weapons have recently begun to receive some attention. Driven by the need to resolve the ethical and operational challenges of using AI in weapons that can autonomously engage targets, and the related objective of potentially establishing regulations and ethical guidelines in this area, the issues on LAWS have been at the forefront. In contrast, the increased attention on the intersection of AI and nuclear weapons (AI–nuclear intersection) is largely driven by states' ongoing nuclear modernisation efforts to ensure operational efficiency. These efforts are necessitated by ageing nuclear infrastructure and the desire to reap benefits from technological innovations or to avoid falling behind adversaries. In this context, China, France, Russia, the United Kingdom and the United States—the five nuclear weapon states (NWS) as defined by the 1968 Treaty on the Non-Proliferation of Nuclear Weapons (NPT)—are seeking to harness AI technology and leverage it in the nuclear domain. As a result, these states are considering the integration of AI into their nuclear operations, including functions that might directly or indirectly affect nuclear decision-making.

A number of possible integrations across the nuclear command, control and communications (NC3) architecture and in systems feeding into it are probably being considered. Although the NWS seem to agree implicitly that nuclear decision-making cannot be fully autonomous and must ultimately rest with human operators, they envision several ways AI can support human decision-makers. However, this raises at least three important concerns. First, not all of the NWS have explicitly declared that humans should have the final say in nuclear decisions and, even if they all did, there is no simple way to verify this, leaving room for grave consequences due to misunderstandings of countries' intentions or AI failures. Second, current deep-learning-based AI models (such as Large Language Models) have specific technological attributes that render them unfit for high-stakes military domains such as the nuclear domain. Third, significant implications arise from the interaction between humans and machines due to human operators placing either too much or too little trust in AI outputs, potentially skewing decision-making even in the absence of AI failures.

These concerns are further exacerbated by the inherent complexity of assessing AI implications within the nuclear context for at least five reasons. Firstly, while some open-source documents from official sources are available on the NC3 systems used by the NWS, most information remains classified due to the topic's sensitivity, allowing only for an approximate understanding of NC3. Adding to the information gap, NC3 systems vary across NWS, as they are tailored to reflect specific capabilities and doctrines.

Secondly, nuclear implications can arise even in the absence of direct AI integration into NC3 components. Adjacent systems that support the NC3 architecture can impact escalation dynamics, indirectly influencing nuclear outcomes.

Thirdly, states may integrate AI into their nuclear enterprises to address different needs driven by unique doctrines and capabilities. For instance, some states may view AI as a tool to compensate for

gaps or inferiorities in specific strategic capabilities. Consequently, potential areas of AI integration will likely differ across NWS, leading to varied interpretations of what could constitute a "strategic advantage".

Fourthly, not all AI applications are potentially risky; they may range from high risk to potentially beneficial, such as those used for training purposes.

Finally, risks are not fully understood: as the technology advances rapidly, it is conceivable that some limitations will be resolved, but new risks might also emerge that cannot be predicted because research has only gone so far. In the aggregate, these elements create significant obstacles for governance.

Based on current AI research, assessing AI implications in specific NC3 functions is not straightforward. It depends on at least three key factors: (a) the specific characteristics (and limitations) of models considered for integration; (b) the specific area where AI will be integrated (in systems within or adjacent to NC3); and (c) the level of human control and redundancies in the automated function. As a result, such an assessment is exceptionally nuanced. Thus, a better understanding of AI implications in the context of nuclear risks and escalation pathways is necessary.

This brief will first introduce the concept of AI, explaining the most widely used techniques and types and differentiating between prior AI techniques already incorporated into NC3 systems. It will then explore the intersection of AI and nuclear decision-making systems, outline possible applications within NC3, and elaborate on the risks and benefits of integration. Finally, it will explore the existing forums for discussion and progress to date, concluding with possible steps forward that could be implemented in relevant forums.

Based on current AI research, assessing AI implications in specific NC3 functions is not straightforward. It depends on at least three key factors... As a result, such an assessment is exceptionally nuanced.

2. The technology

AI encompasses a wide range of methods where machines mimic the way humans think, using highly varied approaches. It is necessary to draw a firm line between rule-based AI, basic machine-learning techniques and advanced techniques such as those based on deep learning, as these present very different risk profiles.

The advanced AI models, which have been at the forefront of public perception with the advent of chatbots such as ChatGPT, differ significantly from the type of rule-based AI that has been incorporated into NC3 since the Cold War. Rule-based AI is used to determine appropriate actions given specific settings. As a result, it performs well with predictable inputs and outputs but is unreliable in complex and uncertain situations, especially those outside its predefined rules.[1] In the context of nuclear command and control, prior applications of rule-based AI during the Cold War included logistical planning related to launch orders and for missile targeting and guidance. Early-warning systems also incorporated a certain level of automation. In this context, AI's role was to provide information to humans in the chain of command, who were then responsible for assessing potential nuclear attacks.[2]

As AI advanced, the advent of machine learning was a breakthrough in that it allowed machines to 'learn' correlations from training data without specific instructions and, therefore, without the need to input rules manually. However, early machine-learning techniques were limited to a narrow set of problems due to their difficulty in generalising and performing multiple functions. Machine learning encompasses a wide range of techniques, including the latest wave of AI spurred by deep learning.

Most recent advances in AI have come from deep learning. Deep learning replicates the way that neurons work in the brain, enabling models to perform complex calculations through layers of artificial neurons. Deep learning-based models, such as large language models (like ChatGPT), have demonstrated an exceptional ability to generalise across diverse tasks and improve continuously with larger data sets and more computational power. These models present an opportunity to enhance military operations by providing faster and more comprehensive data processing from a wide array of sources. Yet these advances also bring notable shortcomings: the reliability and robustness of current AI technologies are not yet sufficient to ensure dependable performance in critical military operations due to their vulnerability to rapid failures.[3]

The reliability and robustness of current AI technologies are not yet sufficient to ensure dependable performance in critical military operations due to their vulnerability to rapid failures.

Indeed, advanced AI capabilities present several attributes that hamper their applicability to high-stakes military platforms, especially those related to nuclear decision-making. At least four key limitations currently exist: unreliability, opacity, susceptibility to cyber threats, and misalignment.

Unreliability

Deep learning-based models can suffer from so-called hallucinations, meaning they can confidently produce incorrect outputs unsupported by their training data. This can mean anything from a chatbot making up facts about a historical event to a vision model 'seeing' things that are not there.[4] In the latter example, AI

can incorrectly identify an object in an image, leading to inaccurate assessments or false positives in critical areas such as threat detection and surveillance.[5]

Opacity

Advanced AI systems function as 'black boxes', meaning it is difficult to understand the underlying processes that lead to an output. As these models learn correlations without specific instructions from humans, it is hard to understand the processes they use to make such correlations. This complexity arises because state-of-the-art deep learning models, such as large language models, can contain billions to trillions of parameters distributed across numerous layers, which are adjusted as the model learns on massive amounts of data. As the models' ability to make good predictions increases, interpreting the way they make these predictions is very difficult. Apart from some limited aspects, we cannot understand how a model goes from the input to the output.

This lack of transparency complicates the verification of AI-generated predictions in critical decision-making scenarios, particularly under tight time pressure in nuclear decisions. However, it is important to note that techniques exist to make this reasoning process transparent or "interpretable", such as mechanistic interpretability, but this results in a trade-off in performance.[6] In practice, this means that advanced AI models tend to fall into two categories that are inversely related: as models become more complex and perform better, they become less transparent; conversely, if they are designed to be transparent (and do not act as black boxes), their performance tends to suffer. Currently, no technique can make large and complex models interpretable without sacrificing some degree of performance.

Susceptibility to cyber threats

AI systems are particularly susceptible to cybersecurity threats in ways that traditional platforms are not, which can open up new avenues for hackers to infiltrate and tamper with sensitive military information. These vulnerabilities provide adversaries and non-state actors with opportunities to compromise AI systems. Concurrently, defensive measures against such cyber threats are inadequate, potentially allowing adversaries to exploit these vulnerabilities in military systems.

Misalignment

As advanced AI models become more and more capable, ensuring they align with human values becomes increasingly critical but remains challenging. Misalignments can lead to grave errors, such as escalating conflicts to nuclear warfare under the guise of pursuing peace. For example, a recent simulation involving five AI models demonstrated their tendency to escalate war, with one model rationalising its move towards nuclear conflict by claiming, 'I just want to have peace in the world'.[7]

3. The intersection of AI and NC3

Assessing the intersection of AI with NC3 is no easy task: open-source documents from official sources on the prospects for AI in the nuclear domain are scarce. This scarcity is compounded by the sensitivity surrounding NC3 systems and the evolving role of AI in nuclear systems based on advances in technology. While informed guesses can be made, some speculation is inevitable due to the nature of the subject and the forward-looking aspect of the discussion.

Despite the limited availability of open-source documents, assumptions can be made about where states might see the best value in AI based on current nuclear postures and on the need to update NC3 systems for operational efficiency. The state with the most transparency on this topic is the US, but even so, no specific official sources tie the role of AI to the nuclear domain, although some sources explore the role of AI in the broader defence domain. One document worth noting is a working paper submitted by France, the UK, and the US at the 2020 NPT Review Conference, which highlights their commitment to preserving human oversight and involvement 'for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment'.[9]

Similar language was replicated in the Responsible AI in the Military Domain Summit (REAIM) Blueprint for Action, a non-binding document reflecting the outcome of the 2024 REAIM Summit, as well as in the original version of the US Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, launched after the first REAIM Summit in February 2023.[10] More recently, on 16 November 2024, US and Chinese leaders jointly affirmed "the need to maintain human control over the decision to use nuclear weapons".[11] Despite the absence of similar statements from Russia, there is a consensus among experts in Russia that human judgement should and will remain central to decisions on nuclear weapon use.[12]

In a recent statement, US Air Force Gen. Anthony James Cotton, commander of the US Strategic Command (STRATCOM), acknowledged the consideration of "all possible technologies, techniques, and methods" for modernising NC3 systems. Within NC3, he noted that AI could enhance decision-making capabilities by automating data collection and processing and speeding up data sharing and integration with allies. At the same time, Gen. Cotton underlined the necessity to keep a human in the loop.[13] Thus, there seems to be a consensus among NWS in applying AI to certain functions such as for intelligence collection and situational awareness tasks, for automating the identification of objects and sensor guidance, and for decision-support roles such as generating real-time operational pictures from multiple sensors.

In these contexts, AI offers the prospect of speed and efficiency by further automating the process of vetting potential missile launches before informing military and political leaders, especially given the growing volume of sensor data. It can also identify pre-launch activities through advanced satellite imagery analysis and potentially discern between different types of attack for more accurate threat assessments. Moreover, AI is seen as particularly valuable for evaluating courses of action in response to potential threats detected.[14]

Overall, AI appears to be most beneficial in functions that are narrow in scope and have redundancy and oversight by design.

Benefits and risks

Overall, AI appears to be most beneficial in functions that are narrow in scope and have redundancy and oversight by design. Employing redundant systems alongside AI can significantly enhance its reliability and safety, ensuring that, in case of a system failure, the overall system is not compromised and can still function correctly.

Certain limitations of AI, particularly hallucinations, could be advantageous in training and war gaming. This would allow military personnel and decision-makers to test out different tactics in simulations presenting unique, unpredictable scenarios. While not always realistic, these scenarios could assist in planning for various potential situations and help personnel become more versatile and better prepared for whatever they might face in actual operations.

However, AI integration presents inherent risks due to the four key technological limitations mentioned above. For example, in decision-support functions, it may be difficult for human operators to understand why AI recommends a particular action due to its black-box nature. This challenge is compounded by AI's tendency to hallucinate, potentially leading to incorrect identification of signals as missile threats or failure to detect actual threats.

Importantly, the adoption of AI technologies by one NWS might trigger a security dilemma for other states. They may feel compelled to either match this technological progress, find asymmetrical responses or revise their military doctrines to negate the perceived advantages and risks of their rival's AI advancements.[15] For instance, significantly advanced AI capabilities that detect enemy movements with unprecedented speed and accuracy might prompt adversaries to develop counter-AI technologies or enhance their cyber warfare capabilities to disrupt or deceive AI systems. As suggested by the previous example, the security dilemma is not confined to AI alone; rather, new technological developments (such as in the context of cyber capabilities or space-based weapons) could lead to a cycle of action and reaction, where states continuously strive to outdo each other to gain strategic advantages, leading to an arms race dynamic.

Finally, issues arise from the interaction between humans and machines. AI systems may reflect the biases of their creators, biasing outcomes, or decision makers may become overconfident (or underconfident) in AI predictions. The rapid pace at which AI operates might also diminish the role of human oversight, turning operators into mere observers of AI-driven decisions.[16] If AI systems appear to possess superior information or make decisions faster than humans can manage, maintaining meaningful human control could become impractical.[17]

New technological developments could lead to a cycle of action and reaction, where states continuously strive to outdo each other to gain strategic advantages, leading to an arms race dynamic.

Further considerations

Although it is possible to categorise the implications of AI integration in NC3 from a strategic stability perspective and by way of technological limitations, the landscape of current and future issues related to this integration is very complex and spans various

interconnected areas. While existing literature on the AI–nuclear intersection does not yet address these issues, it is important to highlight that the nature of any potential risks and benefits of AI integration in NC3 can relate to at least three elements: (a) technological attributes, including vulnerabilities, robustness, reliability, capability and efficiency; (b) the scope of AI's role within NC3 systems affecting operational areas; and (c) the levels of human control and redundancies over automated functions. Many risks and benefits are heavily interconnected as technological attributes directly affect how AI functions in NC3 operations, which in turn affects states' perceptions as well as the countermeasures they might employ, and ultimately, the balance of these elements determines how deterrence calculations shift. In other words, assessing what a "safe" integration looks like depends upon different factors and is not an easy task to determine.

For example, even seemingly beneficial AI models can generate disproportionate risks if deployed improperly: a black-box vision model without verification and redundancy or with vulnerability to hallucinations and cyber threats would result in high levels of risk if integrated into systems related to intelligence collection or early-warning systems. Alternatively, if cybersecurity and hallucination risks can be largely mitigated, the use of such a system in a redundant manner could be beneficial. The critical threshold is that a failure of AI should never result in catastrophic consequences.

However, assessing whether an AI model falls below this critical threshold is further complicated by the fact that nuclear decision-making can be affected even if AI is not directly integrated into NC3 functions. The integration of AI into systems outside the NC3 architecture, such as some intelligence platforms and the conventional domain more broadly, can still significantly impact nuclear decisions. In such cases, potential AI malfunctions or adversarial attacks aimed at data manipulation could spill over into NC3 systems and ultimately influence nuclear decision-making. Although this falls outside the scope of this brief, similar risks may exist in areas such as arms control verification, where incorrect or manipulated data sets could affect escalation dynamics, such as by leading to misinterpretations of compliance or violations.

Existing AI models thus present numerous risks, and the ability to mitigate these risks is currently inadequate. Looking ahead, as technology develops, these capabilities are poised to change, potentially solving some current problems but also generating new ones that cannot be predicted at this point in time. Given these complexities and challenges, it is essential to establish thresholds for AI integration in systems that impact nuclear decision-making. These thresholds can be identified through a risk assessment framework that evaluates how the interaction of the three key variables mentioned above—(a) technological attributes, (b) the scope of AI's role within and adjacent to NC3 systems, and (c) the level of human control and redundancies—can be used to quantify the associated risks.

As technology develops, these capabilities are poised to change, potentially solving some current problems but also generating new ones that cannot be predicted at this point in time. Given these complexities and challenges, it is essential to establish thresholds for AI integration.

4. Forums for debate

There is growing momentum around addressing the intersection of AI and the military domain, exemplified by several initiatives at the governmental level. However, at the time of writing, no current initiative or forum specifically addresses AI in the nuclear domain as a dedicated subject. Despite this, several noteworthy forums and multilateral initiatives discuss AI in the military context more broadly. Although these forums only started to emerge in 2023 and discussions are therefore at an early stage, they provide invaluable platforms where the conversation on military AI is starting to take shape and could eventually incorporate the nuclear angle. This means that they are worth tracking and participating in. These forums include the following:

- **Responsible AI in the Military Domain (REAIM) Summit.** This platform brings together stakeholders, including government officials and civil society representatives, to discuss the opportunities and risks associated with military applications of AI. The first summit took place in The Hague, the Netherlands, on 15–16 February 2023, and the second summit was held in Seoul, Republic of Korea (South Korea), on 9–10 September 2024. The outcome document of this second summit, the "Blueprint for Action", included a key paragraph stating: "it is especially crucial to maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment, without prejudice to the ultimate goal of a world free of nuclear weapons". Among nuclear-armed states, the document was supported by the US, UK, France, and Pakistan. While participating in the summit and the Ministerial-level dialogue, China ultimately decided not to sign the Blueprint.[18]

- **US political declaration on responsible military use of AI.** Launched at the 2023 REAIM Summit, this declaration aims to build international consensus on the safe development, deployment and use of AI in the military. In November 2023, the declaration was revised, consolidating the original 12 principles into 10. New elements were added to address issues arising from human-AI interaction, but the provision on human oversight of nuclear employment was removed. According to confidential sources from US government officials, this decision was reportedly made to accommodate new endorsing states, particularly from the Global South and other parties to the Treaty on the Prohibition of Nuclear Weapons, who expressed concerns that language on nuclear employment could be seen as legitimising nuclear weapons, rather than reflecting any shift in the US position on the matter. As of 10 September 2024, the declaration had been endorsed by 55 states. On 19–20 March 2024, the US held the first plenary meeting with endorsing states to exchange best practices and discuss ways to implement the declaration.

Other venues include the "Capturing Technology - Rethinking Arms Control" conference series, the AI Safety Summits and other informal initiatives. Sponsored by the German Federal Foreign Office, the "Capturing Technology - Rethinking Arms Control" conference series brings together international experts, officials and diplomats to discuss the impact of emerging technologies on arms control. The third conference in this series, held on 28 June 2024 in Berlin, focused on the implications of AI in relation to weapons of mass destruction, including nuclear weapons.

Although these forums only started to emerge in 2023 and discussions are therefore at an early stage, they provide an invaluable platform where the conversation on military AI is starting to take shape and could eventually incorporate the nuclear angle.

One panel was specifically dedicated to exploring the AI-nuclear intersection.

The AI Safety Summit (the first of which was held in the UK in November 2023) offer valuable discussions on the safety risks posed by advanced AI models—although they do not focus on military AI applications. Nevertheless, these discussions may still impact the military debate in other forums. The Bletchley Declaration is particularly important, launched on 1 November 2023 at the AI Safety Summit in the UK. This declaration recognised the safety risks posed by frontier AI models and was signed by, among others, China, France, the UK, the US and the European Union (EU).

The Seoul Declaration, launched on 21 May 2024 at the AI Safety Summit in South Korea, aims to enhance international cooperation on AI governance. A ministerial statement followed, with 27 states and the EU agreeing to collaborate on defining AI risk thresholds. However, unlike the Bletchley Declaration, China refrained from signing the Seoul ministerial statement.

In a similar vein, on the sidelines of the Asia-Pacific Economic Cooperation forum in San Francisco, US, in November 2023, US President Joe Biden and his Chinese counterpart, Xi Jinping, reiterated the need to address AI risks and safety issues, which culminated in the 16 November joint declaration on maintaining human control over the use of nuclear weapons. Earlier, on 14 May 2024, delegations from China and the US met in Geneva, Switzerland, to exchange perspectives on AI safety and risk management. However, it is unclear whether and how these discussions will continue.[19] This bilateral engagement could potentially represent an ideal venue for discussing the risks that AI poses in nuclear decision-making systems. Such discussions could go beyond the current commitments to human oversight in nuclear employment decisions, which alone are insufficient to comprehensively mitigate the complex risks stemming from AI integration. With two major powers engaged in technological competition, this forum offers a critical opportunity to tackle AI safety challenges within the nuclear domain.

Additionally, subgroup two of the Creating an Environment for Nuclear Disarmament (CEND), a US-led initiative aimed at advancing nuclear disarmament, has begun discussions on AI integration into nuclear decision-making systems. This forum provides an interesting platform for discussion, particularly due to the possibility of tackling this issue from a disarmament perspective, such as by exploring the role of AI for arms control and disarmament verification. However, it is still unclear whether discussions on this specific topic will continue and what direction they will take. It is important to note that CEND is a relatively informal initiative with varying levels of state engagement. Despite this, the insights gained from CEND discussions could significantly inform more formal settings.

This bilateral engagement could potentially represent an ideal venue for discussing the risks that AI poses in nuclear decision-making systems.

When it comes to the NPT, AI and other emerging technologies have not been part of the agenda. Although the draft final document of the 2020 NPT Review Conference stated that emerging technologies can affect the risks of nuclear use and can potentially be a challenge for nuclear disarmament, no significant discussion on the AI–nuclear intersection has so far taken place.[20]

As the discussion on AI in nuclear systems is still emerging and the impact on nuclear decision-making remains unclear, significant work is required, particularly in light of the current tense geopolitical environment and widespread perceptions of increasing nuclear risks. The EU could potentially take the following actions:

- **The EU could lead the discussion of the AI-nuclear intersection.** As mentioned, no current forum addresses this intersection as a dedicated subject, presenting an opportunity for the EU to spearhead this critical debate. The EU, which in 2024 implemented the world's first comprehensive AI law, is well positioned to lead such conversations.[21] For instance, in preparation for the next REAIM Summit, the EU could consider creating an AI-nuclear task force to explore potential nuclear risks arising from AI integration in the military domain and incorporate these findings into the REAIM discussions. A critical step would be to engage the NWS and, ultimately, the other nuclear-armed states (India, Israel, the Democratic People's Republic of Korea and Pakistan) in recognising the risks posed by advanced AI in the nuclear domain. Acknowledging that some risks could be catastrophic and lead to nuclear escalation is essential for initiating a meaningful dialogue on mitigating these risks.

- **The EU could call for the inclusion of AI into NPT discussions.** Although the NPT has not yet addressed AI, it should be included in future agendas. States view AI as a strategic advantage, which could potentially increase reliance on nuclear weapons and undermine the treaty's disarmament pillar. Additionally, AI could impact the other two pillars (non-proliferation and peaceful uses of nuclear energy), particularly in the context of non-proliferation and treaty verification. The EU could lead this effort by drafting a working paper for the ongoing review cycle, targeting the 2026 Review Conference and the 2025 Preparatory Committee. Moreover, the EU could utilise unofficial venues to inform these discussions by organising events on the sidelines of future Preparatory Committees and Review Conferences. For example, the US Department of State organised a side event during the 2023 Preparatory Committee on the implications of emerging technologies for future arms control and disarmament agreements. More recently, Germany hosted two side events at the 2024 Preparatory Committee specifically to discuss the impact of AI and EDTs – respectively – on nuclear decision-making. These provided a valuable opportunity to engage NPT delegates in an informal setting while involving all stakeholders, including non-nuclear weapon states (NNWS). Given the high-stakes of AI integration in the nuclear domain, NNWS should undoubtedly be included in this debate.

- **The EU could commission research to better understand the implications of AI in the nuclear domain.** As Gen. Cotton emphasised, there is a need to "direct research efforts to understand the risks of cascading effects of AI models, emergent and unexpected behaviours, and indirect integration of AI into nuclear decision-making processes".[22] Even with limited open-source data on the specific role AI may play in the nuclear systems of NWS, research can still be conducted to methodically assess how different AI models might impact various areas of integration within or adjacent to NC3 systems. By identifying potential nuclear escalation pathways resulting

Thresholds should be based on the principle that any AI failure must not lead to miscalculations or increase the risk of catastrophic outcomes. This research could provide a foundation for developing agreements among NWS to establish risk thresholds.

from AI integration, it is possible to categorise risks and establish thresholds for high-risk applications. These thresholds should be based on the principle that any AI failure must not lead to miscalculations or increase the risk of catastrophic outcomes. This research could provide a foundation for developing agreements among NWS to establish risk thresholds.

References

1 Horowitz, M. C., Scharre, P. and Velez-Green, A., 'A stable nuclear future? The impact of autonomous systems and artificial intelligence', arXiv.org, 13 Dec. 2019.

2 Boulanin, V. et al., Artificial Intelligence, Strategic Stability and Nuclear Risk (SIPRI: Stockholm, June 2020).

3 Hoffman, W. and Kim, H. M., Reducing the Risks of Artificial Intelligence for Military Decision Advantage, Policy Brief (Center for Security and Emerging Technology: Washington, DC, Mar. 2023).

4 A computer vision model refers to an AI system designed to process visual information, such as images or videos. By their very nature, these models could be applied in areas such as surveillance and threat detection. More advanced vision language models and large multimodal models are the latest developments in this field, capable of understanding and generating detailed descriptions from visual inputs.

5 It's important to note that hallucinations in AI models don't necessarily occur due to system malfunctions or errors, but rather because today's advanced AI systems are fundamentally statistical models. In the case of large language models (LLMs), they generate responses based on statistical relationships between words. However, this doesn't fully capture the complexities and nuances of the real world, as the real world doesn't follow the smooth probability distributions that LLMs learn from their training data. As a result, these models are not well-suited for certain critical applications, including those that could impact nuclear decision-making.

6 Mechanistic interpretability is a promising and emerging field that seeks to address the black-box problem by reverse-engineering neural networks to understand the internal reasoning processes that lead to their outputs.

7 It is important to note that these were models tested 'out of the box'. It is likely that these models could be trained to not behave this way as a default, although it is difficult to predict how models will act when they encounter edge cases and things outside their training data. For further detail on the simulation see Rivera, J.-P. et al., 'Escalation risks from language models in military and diplomatic decision-making', arXiv.org, 7 Jan. 2024. For further detail on AI technological limitations in the context of NC3 see e.g. Saltini, A., AI and Nuclear Command, Control and Communications: P5 Perspectives (European Leadership Network: London, Nov. 2023).

8 Examples of official sources that envision the role of military AI include the following: British Ministry of Defence, 'Defence artificial intelligence strategy', Policy Paper, 15 June 2022; French Ministry of the Armed Forces (MAF), L'intelligence Artificielle au Service de la Défense [Artificial Intelligence in Support of Defence] (MAF: Paris, Sep. 2019); US Department of Defense (DOD), Data, Analytics, and Artificial Intelligence Adoption Strategy: Accelerating Decision Advantage (DOD: Washington, DC, June 2023).

9 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, 'Principles and responsible practices for nuclear weapon states', Working paper submitted by France, the United Kingdom and the United States, NPT/CONF.2020/WP.70, 29 July 2022.

10 The Political Declaration was revised in November 2023, and the section addressing human involvement in nuclear decision-making was removed. According to confidential sources from US government officials, this change was reportedly made to accommodate new endorsing states, particularly from the Global South and other parties to the Treaty on the Prohibition of Nuclear Weapons, who expressed concerns over the inclusion of language related to nuclear employment.

11 'Readout of President Joe Biden's Meeting with President Xi Jinping of the People's Republic of China', the White House, statement, 16 November 2024, https://www.whitehouse.gov/briefing-room/statements-releases/2024/11/16/readout-of-president-joe-bidens-meeting-with-president-xi-jinping-of-the-peoples-republic-of-china-3/.

12 Although Russian official sources do not clearly specify their areas of interest for integration, consensus among researchers—including Russian military experts—along with indirect hints from official documents, suggest a shared direction in this regard. For more information see e.g. 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).

13 Hadley, G. 'AI 'Will Enhance' Nuclear Command and Control, Says STRATCOM Boss', Air and Space Forces Magazine, 28 Oct. 2024.

14 Saltini (note 5).

15 Boulanin et al. (note 2).

16 E.g. Israel's autonomous targeting AI system known as 'Lavender' was reportedly designed to identify suspected operatives of Hamas. In this system, human personnel reportedly served only as a 'rubber stamp' for the AI's decisions. For further detail see Abraham, Y., '"Lavender": The AI machine directing Israel's bombing spree in Gaza', +972 Magazine, 3 Apr. 2024.

17 Rautenbach, P., 'Artificial intelligence and nuclear command, control, & communications: The risks of integration', Effective Altruism Forum, 18 Nov. 2022.

18 Brianna Rosen, 'From Principles to Action: Charting a Path for Military AI Governance', Carnegie Council for Ethics in International Affairs, 12 September 2024, https://www.carnegiecouncil.org/media/article/principles-action-military-ai-governance.

19 White House, 'Statement from NSC spokesperson Adrienne Watson on the US–PRC talks on AI risk and safety', 15 May 2024.

20 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, Working Paper of the President on the Final Document, NPT/CONF.2020/WP.77, 26 Aug. 2022.

21 European Parliament, 'EU AI Act: First regulation on artificial intelligence', 8 June 2023.

22 Hadley, G. 'AI 'Will Enhance' Nuclear Command and Control, Says STRATCOM Boss', Air and Space Forces Magazine, 28 Oct. 2024.

The European Leadership Network (ELN) is an independent, non-partisan, pan-European network of over 450 past, present and future European leaders working to provide practical real-world solutions to political and security challenges.

Acknowledgements

This document has been produced with the financial assistance of the EU. The contents are the sole responsibility of the EU Non-Proliferation and Disarmament Consortium promoting the European Network of Independent Non-proliferation and Disarmament Think Tanks, and can in no way be taken to reflect the views of the EU.

This paper was originally prepared as a policy brief for an ad-hoc EU Non-Proliferation and Disarmament Consortium seminar and is being republished with the permission of the European External Action Service (EEAS).

Contact

For further information on the ideas in this report, please contact secretariat@europeanleadershipnetwork.org

Published by the European Leadership Network, February 2025

European Leadership Network (ELN)
8 St James's Square
London, UK, SE1Y 4JU

@theELN | europeanleadershipnetwork.org

EUROPEAN LEADERSHIP NETWORK

European Leadership Network
8 St James's Square
London, SE1Y 4JU
United Kingdom

Email: secretariat@europeanleadershipnetwork.org
Tel: 0203 176 2555

Follow us

europeanleadershipnetwork.org