EUROPEAN LEADERSHIP NETWORK

AI and nuclear command, control and communications: P5 perspectives

Alice Saltini

November 2023

The European Leadership Network (ELN) is an independent, non-partisan, pan-European NGO with a network of over 300 past, present and future European leaders working to provide practical real-world solutions to political and security challenges.

This research was performed with the generous support of U.S. Department of State's Bureau of Arms Control, Verification and Compliance.

The author would like to give special thanks to Oliver Meier, Rishi Paul, and Graham Stacey for their guidance throughout this project. Thanks is also due to Megan Karlshoej-Pedersen and Edan Simpson for their copyediting support, and Esther Kersley for her design support, as well as the authors of the four bibliographies for their outstanding contributions to this project. Thanks also to Jane Kinninmont, Wyatt Hoffman, Wilfred Wan and project participants for providing valuable feedback during the early drafts.

About the author

Alice Saltini
Research coordinator for the European Leadership Network

As the Research Coordinator at the European Leadership Network (ELN), Alice Saltini is actively involved in a range of projects within the Global Security Program that include an examination between the interplay of AI and nuclear risks. With a keen interest in AI regulation at the intersection with nuclear systems, Alice is dedicated to developing policy solutions in this field, contributing constructively to ongoing discussions.

Believing in the importance of transparency in AI, Alice seeks to understand and address challenges presented by the potential implications of advanced AI models for nuclear command and control systems. Her insights have found a platform in a number of multilateral fora, such as the NPT.

Prior to joining the ELN, Alice interned for the Comprehensive Nuclear-Test-Ban Treaty Organization in the External Relations, Protocol and International Cooperation Section, and worked as a Research Assistant at the James Martin Center for Nonproliferation Studies. Alice is also a recent cohort of the CTBTO Youth Group's CTBTO-CENESS Research Fellowship 2022.

She holds a Master's degree in Russian studies and a Post Graduate Certificate (Pg Cert) in Nonproliferation Studies from the Middlebury Institute of International Studies, benefiting from a full merit-based scholarship during her tenure at the Middlebury Institute.

Combining her academic background with practical experience, Alice hopes to foster informed dialogues about the integration of AI in the nuclear domain.

Contents

Executive summary

The nuclear-weapons states China, France, Russia, the United Kingdom and the United States are increasingly recognising the implications of integrating artificial intelligence (AI) into nuclear weapons command, control, and communication (NC3) systems. Although the risks and pitfalls are widely acknowledged, the allure of gaining a strategic edge in a rapidly changing nuclear landscape, combined with fears of lagging behind competitors' AI innovations, might push these states in a race to integrate AI technologies into NC3 systems whose reliability is yet to be verified.

AI is not a new term, and techniques falling under the umbrella are already integrated into NC3 systems in nuclear-weapons states. But the possible integration of today's state-of-the art deep learning-based AI models presents a significantly different and larger set of challenges than that of current rule-based models. Significant concerns exist regarding the reliability and suitability of powerful large models for critical functions such as affecting nuclear weapons decision-making.

The project conducted by the European Leadership Network (ELN) titled "Examining the impact of artificial intelligence on strategic stability: European and P5 perspectives", supported by the US Department of State's Bureau of Arms Control, Verification and Compliance, established how nuclear-weapon states are using and seeking to use AI technologies in their NC3 systems and addressed the repercussions of such integration.

Exploring the risks inherent to today's advanced AI systems, the report sheds light on characteristics and risks across different branches of this technology. It establishes the basis for a general-purpose risk assessment framework to analyse models being considered for integration and forms the basis of norms and a threshold for a moratorium of integrating high-risk AI systems into NC3.

During a series of confidential dialogues, experts from the nuclear-weapon states compared their assessments of risks and benefits of using AI in NC3 systems, with a view to establishing guardrails against nuclear escalation in crisis scenarios.

Core findings of the project include:

- **The way in which nuclear-weapons states integrate AI into NC3 systems are different**, reflecting specific nuclear doctrines, military cultures, civil-military relations, and ethical considerations. However, they all see the value in AI for improved situational awareness, early threat detection, and decision-support.
- **All nuclear-weapon states emphasise the importance of human oversight in nuclear decisions.** They agree with the concept of keeping "humans-in-the-loop", although it is unclear to what degree their interpretations overlap.

Exploring the risks inherent to today's advanced AI systems, the report sheds light on characteristics and inherent risks across different branches of this technology.

- **Integration of cutting-edge AI models, such as large language models, poses exceptional risks to strategic and crisis stability** because of their opacity and unreliability.
- **Nuclear-weapon states should agree to swiftly impose a moratorium on the integration of high-risk AI models.** The moratorium can be elaborated on the risk profiling system introduced in this report to provide a scoring system that enables the classification of high-risk AI systems.
- **For AI models that do not carry the same high-level risks, bilateral initiatives at the track-1 level should revolve around the retention of human control over nuclear systems.** Concurrently, track-2 dialogue should delve into technical subjects, such as practical ways to assure human oversight.

Nuclear-weapon states should agree to swiftly impose a moratorium on the integration of high-risk AI models.

Introduction

The integration of artificial intelligence (AI) in military systems, particularly its implications for nuclear command, control, and communication (NC3), is an increasing focus among the P5 states. Nuclear-weapon states have already integrated AI into their NC3 systems. However, these systems have little to do with the predominant AI models of today. The consideration on whether to integrate advanced AI models is driven by their fast-paced evolution and the ongoing modernisation of NC3 systems.

The term AI encompasses a wide range of methods in which machines mimic how humans think. AI has many fields and approaches, which are highly varied and present different levels of risks. However, integration of today's state-of-the-art AI models, fit for a wide range of tasks, into NC3 systems presents the greatest danger to strategic stability. By and large, this stems from advancements in deep learning that led to large language models (LLMs) and today's generative AI. These powerful models have demonstrated tremendous ability when it comes to solving complex tasks, but lack reliability and interpretability, which makes them largely unfit (as is) for integration in critical decision-making and support systems.[1,2]

Official sources provide few details about how the P5 states envision cutting-edge AI models playing into their nuclear development and modernisation programmes. But what is certain is that current advancements within AI technologies will influence nuclear weapons, postures, and decision-making processes.[3] Given the potentially profound implications of AI on military capabilities, nuclear-weapons states (NWS) have made the pursuit of AI a strategic priority. While the US aims to utilise AI to maintain its technological superiority in defence and military applications, Russia associates AI advancements with its global standing and sovereignty.[4] China is actively exploring 'intelligentised' warfare, where AI and advanced technologies are integrated comprehensively across all levels of warfare, as well as the military-civil fusion sector.[5] The UK and France have also marked AI as a cornerstone of their national defence strategies.[6]

Whether these nations' goals are to maintain or achieve technological superiority, as seen with the US, Russia, China, and the UK, or to sustain a certain level of strategic autonomy on AI, as in the case of France[7], the pursuits bring along significant challenges. This is true both when it comes to the technological perspective and broader concern of international peace and security. For one, in its current form, advanced AI models have vulnerabilities that stem not only from inherent technological limitations, but also from human interaction with such tools. Additionally, from a global security perspective, AI introduces risks of miscalculations, misperceptions and misinterpretations, whether from system failures, vulnerabilities, or misuse, potentially leading to inadvertent or accidental escalation.[8]

From a global security perspective, AI introduces risks of miscalculations, misperceptions and misinterpretations, whether from system failures, vulnerabilities, or misuse, potentially leading to inadvertent or accidental escalation.

On the other hand, AI provides benefits such as real-time processing of vast datasets. Such benefits, combined with the pursuit for a military edge and decision advantage (or the desire to prevent adversaries from gaining strategic advantages), as well as the need to modernise legacy NC3 systems to ensure technical efficiency and operational integrity, could potentially prompt P5 states to prioritise swift AI integration. Although some experts argue that the integration of AI into nuclear-weapon systems,

particularly within the NC3 framework, is advancing at a more deliberate pace compared to its integration in other military sectors (due to the inherent unpredictability, susceptibility, and opaqueness of today's complex algorithms), it is crucial to consider the factors that would counteract such a cautious approach.[9] For instance, even as P5 decision-makers appear to understand the perils of deploying untested and unreliable AI capabilities, the appeal of securing an early advantage in an evolving, multipolar nuclear environment, coupled with the fear of potentially falling behind a rival's AI advancements, could exacerbate the security dilemma and therefore spur the adoption of systems whose reliability is yet to be verified.[10]

Despite the many military applications of AI, ranging from intelligence gathering to logistics and training, their possible future integration with nuclear decision-making draws widespread concern. As stated by Izumi Nakamitsu, UN High Representative for Disarmament Affairs;

> "Technological advances and the emergence of new domains in cyber and outer space have exposed new and dangerous vulnerabilities. This is especially the case when it comes to nuclear command and control structures, where the use of technology in and against those structures could again lead to mistake or miscalculation, not least through the intervention of malicious third parties".[11]

To shed light on these pivotal problems, the ELN launched a project titled 'Examining the impact of artificial intelligence on strategic stability: European and P5 perspectives'. The project built from the following pressing questions: What are the repercussions of AI integration into the NC3 framework for nuclear decision-making? How do advanced AI models contrast with current AI in NC3 systems? How do the diverse perspectives and needs of P5 states influence AI integration? How does the perception of AI-related risks vary across P5 states? Is there common foundation from which we can work to explore a future where AI models are integrated to effectively support, rather than supplant, human decisions?

The project aimed to address these questions by focusing on the AI-nuclear intersection from P5 states' perspective. Through this endeavour, the objective was to propose risk mitigation measures for the P5 states, drawing from diverse stakeholders, including industry experts and academics to craft informed and effective strategies.

Research focus and methodology

This research had three primary objectives. First, it aimed to evaluate the potential risks and benefits associated with AI from a technological and global security perspective. Specifically, the research was focused on understanding the ramifications of AI in the context of nuclear weapons decision-making. This required a consideration of how more advanced AI technologies, such as those based on deep learning, could be integrated into NC3 systems, based on ongoing discussions within the expert community.

Second, the goal of the project was to analyse different perspectives on AI risks and benefits based on different needs and applied to specific criteria across P5 states, such as nuclear doctrines and postures. However, AI application by individual countries is constantly evolving and often lacks transparency due to the sensitivity surrounding NC3 systems. Notably, a growing body of literature delves into the potential applications of AI technologies, especially in the context of NC3 modernisation, with a particular focus on the US.[12] However, there is limited publicly available information regarding AI endeavours related to the nuclear-weapon systems of the remaining P5 states.

To address these gaps in the literature and to identify shared perspectives among P5 states, the ELN commissioned four bibliographies from China, France, Russia, and the UK. The primary objective was to analyse these countries' perspectives on AI and its military and nuclear applications. These analyses specifically emphasised the potential impact of AI on nuclear decision-making.

An initial workshop discussed the analysis of these bibliographies. Here, the authors of the bibliographies presented their research findings, shedding light on the internal debates within each of the four countries. The presentation of these bibliographies laid the groundwork for subsequent discussions, which aimed to identify commonalities among the P5 states and to establish a basis for preliminary policy recommendations with a view to developing guardrails for the use of AI in NC3 systems.

Additionally, the project aimed to formulate an effective approach to risk reduction among P5 states. For this purpose, the ELN convened a second workshop to tease out the insights from the first workshop and to formulate policy recommendations for risk reduction, tailored to the P5 states. The workshop brought together experts from all P5 states, and included analysts, scholars, defence practitioners, and representatives from the private sector. The event leveraged two tabletop exercises designed to explore potential escalation or de-escalation pathways in the context of the belief that an adversary had integrated AI into NC3 systems or was employing AI-enhanced tools.

The intersection of AI and NC3

Understanding the technology

AI techniques have been incorporated into NC3 systems since the Cold War era.[13] For example, rule-based systems, which have already been integrated into NC3, excel in tasks with predictable inputs and outcomes, but are less reliable in complex, uncertain situations.[14] However, comparing these systems to the predominant approach used in AI today is not helpful because of the vast differences between today's deep learning and earlier approaches.

AI can be described as computerised processes that emulate tasks traditionally performed by humans. Machine learning algorithms encompass a vast range of techniques that allow machines to 'learn' from data without explicit instructions. In machine learning there are different learning paradigms ranging from unsupervised to supervised to reinforcement learning. There are many intersections between these methods and hybrid approaches.[15]

As of 2023, AI is predominantly associated with deep learning.[16] The advent of generative AI with LLMs has brought this technology to the forefront of public perception. This approach is based on neural networks which loosely mimic the way neurons function in the human brain, facilitating complex calculations through layers of artificial neurons. Deep learning, a form of neural networks, employs an architecture with multiple layers of artificial neurons. This method is termed as 'black box' because its internal decision-making process is opaque and hard to decipher.

Earlier machine learning methods due to their their difficulty in generalising were applied to only a narrow set of problems. The introduction of the transformer architecture marked a significant shift. This architecture demonstrated an unparalleled ability to generalise across diverse tasks and to continuously improve as the size and quality of the training datasets increased. This has led to the development of tremendous capabilities. Nevertheless, these capabilities also bring many unresolved problems that severely hamper their applicability to sensitive domains. These are related to significant unreliability, which manifests itself in two major ways. It tends to generate incorrect outputs and is vulnerable to cyber-attacks. On the former, this can include confidently producing incorrect data unsupported by training data. This unreliability is compounded by the 'black box' nature of AI, making it difficult, if not impossible, to discern the underlying processes leading to a particular output.

If incorporated into strategic decision-making systems, these models pose some of the gravest risks due to their opaque nature, unpredictability, and susceptibility to cyber-attacks.

If incorporated into strategic decision-making systems, these models pose some of the gravest risks due to their opaque nature, unpredictability, and susceptibility to cyber-attacks.

In the specific context of deep learning, recent progress in generative AI has been notable. Generative AI primarily focuses on creating content, distinguishing it from other AI systems designed for tasks like classifying or grouping data.[17] These advancements in generative AI have largely been driven by the ability to massively scale models by increasing the quantity and quality of data and the computation invested.

A prime example of generative AI are LLMs. These models are trained on vast bodies of text, breaking them into units called

tokens. The primary function of these models is to predict subsequent tokens in a sequence. When they are trained on a large enough scale, they learn an underlying representation of the data, allowing them to produce humanlike text. Recent advances have underlined the role of high-quality data when it comes to enhancing the capabilities of such models.[18]

Central to LLMs is the transformer architecture, a subset of deep learning models. Transformers are capable of being trained on massive scale due to their parallelism.[19] They employ an attention mechanism, which allows the model to focus on relevant parts of the input data, akin to how a human reader might pay more attention to key phrases in a document. Transformers spawned the rapid emergence of today's LLMs and large multimodal models. They form the basis of most of these models. There is active development of new architectures, which could result in new models. If a more effective architecture emerges, it might eclipse the role of transformers.[20]

Understanding these models is vital for grasping the technological risks associated with incorporating AI into NC3 systems. Attempts to develop machine learning models which are generally transparent and easily interpretable are known as 'explainable AI' (XAI), or explainable machine learning (XML).[21] In contrast, the inner workings of complex models like deep neural networks remain largely opaque due to the vast number of parameters that are tuned during training.[22] This lack of understanding of the models' workings is particularly concerning in critical areas like nuclear weapons decision-making, where understanding the reasoning behind decisions is crucial for ensuring system safety and maintaining trust in its functions.

There's a divide within the expert community on the suitability of advanced AI for integration in NC3 systems, with some arguing that the unpredictability and complexity of today's AI algorithms make them unsuitable for critical safety applications, particularly given the 'black box' nature of certain machine learning models.[23] Others believe that sophisticated AI technologies like deep learning could offer strategic benefits (such as enhanced real-time data analysis, improved prediction accuracy, and automation of complex tasks), albeit with substantial risks, including digital vulnerabilities and the introduction of new uncertainties.[24]

Understanding the categorisation of AI models is particularly relevant as P5 states are debating the integration of generative AI models in military systems, sparking an intense debate and showcasing the willingness to pioneer the integration of generative AI in defence systems before their adversaries.

Understanding the categorisation of AI models is particularly relevant as P5 states are debating the integration of generative AI models in military systems, sparking an intense debate and showcasing the willingness to pioneer the integration of generative AI in defence systems before their adversaries.[25] It is worth noting that the debate around integrating generative AI does not specifically reference nuclear weapons. This absence doesn't automatically mean that governments are planning (or not planning) to incorporate these models into their NC3 structure. Understanding the technological risks of such systems becomes exceedingly critical for the ongoing debate.

Historical context and current applications

During the Cold War, both the US and the Soviet Union created advanced command and control infrastructure that were

designed to rapidly identify threats and formulate appropriate responses within a short timeframe.[26] For instance, both countries incorporated AI automation for functions including the logistical planning related to launch orders and missile targeting and guidance.[27] Similarly, the US and the Soviet Union worked on automating early-warning systems to some extent, with the goal of rapidly providing critical radar data to decision-makers. In this context, the need for automation was attributed to the need to allow officials more time to deliberate on a retaliatory strike.[28] Historically, these systems primarily fed information to officers in the chain-of-command, who then needed to assess if an opponent initiated or was gearing up for a nuclear launch.[29]

However, both countries soon recognised inherent limitations in automating the practice of nuclear deterrence, particularly due to the proneness towards generating false alarms.[30] As a result, it became clear that human oversight was indispensable for verifying the data generated by these systems, as well as for making the ultimate decisions regarding nuclear launches. Automation in command and control was generally considered viable only in scenarios where human decision-makers would be physically incapacitated and thus unable to assess the situation or make judgments.[31]

AI integration in modern NC3 frameworks, which have been updated with more modern technology and modified to address evolving threats, presents varying risk profiles and could include machine learning algorithms of unknown types. Below is a brief rundown of core NC3 operations where AI might be, or has already been, integrated.

Early warning systems differ across nuclear states and are subject to a limited availability of information. They have three tasks: detection, warning, and attack characterisation.[32] These systems encompass radar, sonar, and infrared detection tools placed in sensors in space, on land, airborne, and both beneath and on the water's surface.[33]

Reliable communication is a crucial component across all aspects of NC3. Whether it's early warnings, strategic communication, executing orders, or force management, robust communication channels are essential. These systems need to not only be secure but also resilient against both physical and cyber threats.[34]

Decision support systems are designed to assist leaders in making nuclear weapons-related decisions, such as the deployment, use, or transportation of weapons. Among other things, they might encompass systems that process signal data, suggest courses of action, provide targeting options, and decide on pre-selected targets based on simulations.[35]

AI integration in modern NC3 frameworks, which have been updated with more modern technology and modified to address evolving threats, presents varying risk profiles and could include machine learning algorithms of unknown types.

Risks and benefits calculus

When it comes to potential benefits, AI offers the prospect of speed and efficiency by expediting the process of human operators to vet potential missile launches before informing military and political leaders. This is especially true as sensor data volumes grow. AI can quickly synthesise data from diverse sources, including military and civilian sensors, and it can offer quicker threat identification and

continuous system health monitoring. This can potentially enhance communication systems by bolstering cybersecurity defences and by improving resilience. It can also identify pre-launch activities through analyses of advanced satellite imagery, potentially discern between different types of attacks to provide more accurate threat assessments, and potentially even track submarines with advanced sea-based sensors.[36] Moreover, the integration of AI into such systems would arguably come with a reduction of human-induced errors, biases, and fatigue-driven mistakes.[37]

In the context of decision support, AI technologies have the potential to enhance information processing and situational awareness. AI can gather and integrate diverse data sources. It can also analyse trends and anomalies at the front lines.[38] The potential of AI to pre-emptively gauge nuclear threats, dubbed 'predictive forecasting', is especially notable: for instance, AI would be able to pre-emptively analyse troop movements, supply lines, and other intelligence to predict threats before they emerge.[39] The allure of AI is further underscored when considering emerging technologies such as hypersonic vehicles. Traditional human cognitive processing speeds simply might not cut it in such scenarios, necessitating AI-augmented systems for timely and effective responses.[40]

However, AI integration presents inherent challenges, due to both technological limitations as well as problems with human-machine interactions. As far as technological limitations are concerned, the 'black box' nature of some AI models, makes its decision-making process very challenging to decipher with our current understanding of these models. If integrated in strategic decision-making systems, AI predictions and decisions would be neither accountable nor verifiable. Situations where AI models mistakenly draw conclusions with unwavering confidence, known as 'hallucinations', are particularly troubling.[41] From an NC3 perspective, a single incorrect output from an AI system could translate to grave misunderstandings, such as mistaking natural phenomena for missile threats. Furthermore, cybersecurity vulnerabilities introduce an added dimension of risk, where a compromised AI system could compromise situational awareness, potentially resulting in grave consequences based on flawed or tampered AI data. This concern could lead to dedicated efforts by state and non-state actors to compromise an adversary's AI systems. When considering the training data for these AI systems, the scarcity of real-world nuclear escalation scenarios necessitates a reliance on synthetic data. This dependency raises concerns regarding data reliability, which can obviously affect NC3 systems.[42]

The 'black box' nature of some AI models, makes its decision-making process very challenging to decipher with our current understanding of these models. If integrated in strategic decision-making systems, this might leave no accountability systems and no method of verification for AI predictions and decisions.

Regarding problems related to human-machine interactions, there's potential for AI models to mirror the biases of their human developers, which can result in skewed conclusions. In this context, humans may either unduly trust ('automation bias') or be overly sceptical ('trust gap') about AI-generated outputs. The speed of AI-driven operations could conceivably replace human operators from their traditional oversight roles, relegating them to passive observers of AI decisions or, in extreme cases, put them entirely 'out of the loop'.[43] Genuine human control might become unrealistic if AI systems hold an informational advantage or operate too quickly for humans to intervene in real-time.[44]

Given the above-mentioned AI risks and possible benefits, from a national security perspective, AI also has the potential to affect strategic stability. As highlighted earlier, the adoption of cutting-edge AI technologies within one nuclear state might trigger a security dilemma, wherein another state might either attempt to match those technological advancements, find asymmetrical responses, or adjust their doctrines to counterbalance the technological edge presumably offered by the technology. This could potentially result in a destabilising effect on the existing balance of power and raise the likelihood of nuclear weapon use.[45]

Moreover, AI has the potential to accelerate data analysis, which, in turn, could create opportunities for de-escalation and thus lessen ambiguity during a crisis.[46] However, the swift decision-making ability of AI can, paradoxically, contribute to rapid and unintended escalations. As warfare gets redefined by AI speeds, there's a tangible fear of humans being left behind, unable to control unfolding scenarios.[47] Moreover, it's important to note that for AI to effectively reduce ambiguity in such critical situations, the integrated AI systems must be both reliable and transparent. Unfortunately, AI systems are susceptible to rapid failure. Although significant efforts are underway to enhance the robustness of AI systems and enable them to at least operate reliably in the presence of adversarial interference, the techniques for improving robustness have not yet reached a level of maturity that can guarantee reliable performance.[48] At the same time, ongoing research aimed at enhancing interpretability of deep learning models has not yielded considerable results.[49]

Additionally, AI's advancements in the domain of conventional weaponry can also affect strategic stability, particularly in the relationship between conventional and nuclear forces. The application of AI to non-nuclear strategic weapons or the creation of novel weapon systems might threaten the survivability of nuclear assets. These AI-augmented conventional capabilities can have a destabilising effect, especially on nuclear-armed states that are technologically weaker or unable to match these AI-driven advancements, potentially driving them to further develop or modernise their nuclear capabilities.[50] Given that this may challenge the survivability of their retaliatory-strike capability, this may lead nations to adopt a launch-on-warning posture, which entails launching a retaliatory nuclear counterstrike upon receiving a signal of an impending adversary nuclear attack.[51]

Risk determination

This report has so far examined AI-generated risks from the vantage points of the P5 states and assessed the impact of such technologies on strategic stability. However, determining AI-generated risks and benefits for each possible application within the NC3 framework is inherently challenging for several reasons.

The risk profiles of AI models are not uniform. Within deep learning alone, vast differences exist both among architectures and individual models.

Firstly, the risk profiles of AI models are not uniform. Within deep learning alone, vast differences exist both among architectures and individual models. Moreover, the area of AI application also influences its associated risks. For example, deep learning-based models can be trained to perform exceptionally well in complex tasks, but they do so without revealing their internal decision-making processes. This lack of understanding can have different

implications based on the NC3 application. For instance, the hypothetical integration of these models would have a lower impact on applications such as communication path optimisation, as opposed to integration into strategic decision-making systems, where the need to understand AI-generated outputs arguably matters far more.

This is particularly relevant because deep learning-based models offer groundbreaking potentials but, at the same time, these models lack reliability because of their inherent 'black box' nature, making their decision-making processes challenging to decipher.

Another dimension which impacts risk assessments pertains to the context and manner in which AI systems are deployed. The impact of a system glitch can range from minimal in certain areas, like communication path optimisation, to very high in realms directly linked to strategic decision-making. The spectrum of autonomy, spanning from light to complete (based on the level of human supervision required), further complicates this assessment.

To navigate these complexities, this report proposes a framework for evaluating the risks of AI integration within NC3 systems, along with the creation of metrics for risk profiling of AI systems.

View from the expert community

Although deep learning architectures are arguably the main way in which second-wave AI will be used in the nuclear domain, including in nuclear command, there appears to be general confusion among the policy community on what exactly AI is, and how its various subsets might affect the NC3 systems differently.[52] This general confusion within the arms control community was emphasised by participants at the workshop and is also reflected in P5 countries' internal debates. This is further compounded by secrecy surrounding the types of algorithms that P5 states intend to integrate in their NC3 systems.

During the tabletop exercises organised by the ELN, the risks and possible benefits of AI (along with its impact on strategic stability) were explored, and several key observations emerged.

1. **Information overload and subsequent discrepancy in AI's availability and utility.** The introduction of AI augments the decision-making process with an influx of data, which could lead to information overload. During a crisis, leaders may revert to simpler decision-making inputs. Notably, during the exercises, many participants opted for traditional systems over AI, not just for early warning, but also for data analysis.

2. **Confirmation bias.** The trust in AI's recommendations, particularly during crises, seems conditional. Leaders are more likely to trust AI's advice if it aligns with their pre-existing beliefs. For instance, if AI suggests aggressive measures potentially leading to nuclear escalation, leaders already leaning towards aggression might find this recommendation more palatable. On the other hand, those reluctant to escalate tensions may disregard AI suggestions advocating for a confrontational stance.

The research highlighted general confusion among the policy community on what exactly AI is, and how its various subsets might affect the NC3 systems differently.

3. **Automation bias.** The exercises unveiled a predisposition, notably among the younger generation, to trust AI's outputs. This indicates a potential trust disparity in AI based on age, which might stem from varying levels of familiarity with technology.

4. **Prioritising AI interpretability and oversight.** Determining the reliability of AI-enhanced systems often becomes possible only after a crisis has passed. This highlights the need to make AI models more interpretable. Participants frequently emphasised the importance of understanding the logic behind AI decisions. Pairing this with strong human supervision can increase trust in AI-integrated NC3 systems.

5. **Value of independent verification.** Participants emphasised the importance of separate early warning systems that don't solely rely on the same AI data sources, allowing for an additional layer of validation and reducing the chances of misinterpretation or errors.

P5 views on the integration of AI in nuclear decision-making

To identify shared perspectives among P5 states, the ELN commissioned four bibliographies on Chinese, French, Russian, and British perspectives on AI integration in nuclear decision-making, from a range of non-governmental experts. A brief description of each country's position is presented below, including the US perspective for a comprehensive analysis of P5 states.

The following table leverages the ELN-commissioned bibliographies to pinpoint references in the literature on states' interests of AI applicability in the NC3 domain. Although not exhaustive, the table categorises mentions from official and non-official sources where applicable, which are further explored in the bibliographies. The information pertaining to the US is sourced from independent research of open-source documents.

P5 states views on AI applicability

	US	UK	France	China	Russia
Early Warning Systems	Official sources: highlight the US' need to enhance integrated tactical warning and attack assessment.[53] Expert community: highlights AI utility when it comes to improving the speed and precision of early warnings.[54]	Expert community: highlights AI utility for data analysis and prediction, as well as for threat detection.[55]	Official sources: highlight the utility of AI in information processing and in enhancing early warning systems.[56] Expert community: highlights AI as potentially beneficial for early threat detection.[57]	Official sources: stress the importance of improving strategic early warning.[58] Expert community: emphasises the importance of AI for data collection, analysis and threat identification.[59]	Expert community: highlights AI as potentially beneficial for swift threat detection and damage prediction[60,61]
Intelligence Surveillance and Reconnaissance	Official sources: highlight AI utility for better situational awareness.[62] Expert community: highlights AI utility for intelligence collection and analysis, sensor data fusion, as well as anomaly detection.[63]	Official sources: highlight AI utility in data fusion and analysis, as well as detection and identification of objects.[64] Expert community: highlights AI utility for surveillance of dangerous environments and processing real time data.[65]	Official sources: highlight AI utility in optimising cross-referencing of multi-source data for intelligence collection.[66] Expert community: highlights AI utility for better situational awareness.[67]	Expert community: highlights AI utility in the operational context, such as in the prediction of enemies' nuclear weapons deployment.[68]	Expert community: highlights current use of AI for surveillance and protection of stationary and mobile objects, video surveillance, and safeguarding patrol routes.[69]

	US	UK	France	China	Russia
Decision-support functions / Command and control	Official sources: highlight the US' need to optimise resilience approaches for NC3 architecture with advanced decision support technology and for integrated planning and operations.[70] Expert community: highlights AI utility for improving information processing in support of human decision-makers.[71]	Official sources: highlight AI utility for better-informed decision-making and planning support.[72] Expert community: highlights AI utility in predicting and suggesting responses to enemy actions in real time.[73]	Official sources: highlight utility of AI in decision-making for planning and operations, as well as evaluating courses of action[74] or detection of anomalies.[75]	Official sources: emphasise command and decision-making, as major areas of interest for advancing the role of AI in national defense.[76] Expert community: highlights AI utility in strengthening nuclear command and control by providing better situational analysis, combat guidance, and decision-assistance.[77]	Expert community: highlights AI current use in decision-support of day-to-day activities and operational combat management. Moreover, sources highlight AI utility in support of retaliation planning.[78]
Precise Delivery of Nuclear Assets / Targeting	Official sources: highlight AI utility in improving targeting ability.[79] Expert community: highlights AI utility in improving target identification, delivery system reliability, penetration accuracy.[80]	Official sources: highlight AI utility for target detection.[81] Expert community: highlights AI utility to detect, track, target, and intercept missile, air, and space defence systems.[82]	Official sources: highlight AI utility for target detection.[83] Expert community: highlights the potential utility of AI improving strike precision.[84]	Official sources: highlight the utility of AI in improving targeting and missile guidance.[85] Expert community: highlights AI as useful in improving targeting accuracy and weapons delivery.[86]	Expert community: highlights the need for radar systems to tackle new tasks that are poorly performed by traditional AI algorithms, such as target recognition (including for strategic conventional and novel nuclear strike capabilities).[87]
Enhance Secure Communication	Official sources: highlight the need to strengthen NC3 systems by enhancing protection from cyber-attacks and improving communication links by using AI.[88] Expert community: highlights AI utility in accelerating dissemination of orders.	Expert community: highlights AI utility in information dissemination.[89]	Official sources: highlight AI utility when it comes to enhancing communication systems against cyber threats[90] and for resilient communication networks.[91]	Expert community (including from the PLA): highlights AI utility for cyber network defence monitoring and protection.[92]	Expert community: highlights AI utility in enhancing communication channels[93] and potentially for semi-automatic launch order transmission.[94]

China regards military applications of AI as critical for improving its military capabilities and ensuring strategic stability, aspiring to become the global leader in AI by 2030. While details on China's AI warfare advancements in official sources remain limited, Chinese analysts highlight AI's potential benefits in early warning, intelligence, surveillance and reconnaissance (ISR), and nuclear decision support systems. Nevertheless, they also recognize risks associated with AI use in NC3, such as shorter decision-making time, inaccurate or tampered data and, potentially, crisis instability and escalation. Moreover, Chinese analysts have pointed to the current immaturity of AI technologies when it comes to supporting nuclear decision-making.[95] They emphasise that AI-enabled NC3 systems could afford nations a strategic edge, possibly endangering other states' second-strike capabilities and potentially increasing temptations of a pre-emptive strike.[96] Consequently, China states that nuclear-weapons states should refrain from using AI-enabled systems to strike each other's strategic assets, including nuclear capabilities.[97] Chinese official sources highlight the importance of the human role in deciding the use of weapon systems. While they haven't directly addressed AI's role in nuclear weapons systems or clarified whether there is agreement to keep 'humans in the loop' for critical decisions concerning nuclear weapons, an informal consensus among policy experts, academics, and some PLA officers suggests that AI will not supplant humans in strategic decisions. Even with AI's rapid advancements, the final authority should rest with humans, according to Chinese mainstream thinking. Concurrently, China is actively working to enhance its AI capabilities and investing in expert training to boost both its economic strength and military prowess.[98]

France acknowledges the strategic significance of AI in strengthening its military capabilities and in maintaining competitiveness. The country identifies AI as invaluable in various military sectors including force training, data analysis, logistics, and operational support. While official sources recognise AI's transformative power in many military areas, there's a noticeable absence of discussion surrounding its role in the nuclear weapons domain. The reason for this is two-fold. Firstly, current AI technologies are viewed as too immature to be seriously addressed in research or doctrines, and secondly, given a tradition of conservative debate in France about nuclear deterrence, discussion of AI can be viewed as deviation from the balance of nuclear deterrence.[99] France adopts a cautious approach towards integrating AI into decision-making processes, given the inherent risks and ethical concerns.[100] It recognises that the advantage in decision-making for armed forces will depend on their capacity to proficiently handle data for operational ends; and that the evolution of AI tools can reinforce interconnectivity with allies and partners. Official sources highlight the centrality of the role of humans in nuclear decision-making, assuring that they will always play a role. France's primary concern isn't rooted in the use of AI itself. Rather, it's concerned about other nuclear countries excessively relying on AI, or that AI could be misused unethically. Therefore, the integration of AI into nuclear forces and decision-making structures of other countries merits scrutiny due to its possible influence on global strategic balance, according to the French perspective.

Chinese analysts have pointed to the current immaturity of AI technologies when it comes to supporting nuclear decision-making.

Russia perceives the military integration of AI as essential and necessary for contemporary warfare. Notwithstanding understood risks, and in a bid to remain competitive with nations heavily

funding AI to gain military superiority, Russia considers investment in military AI a necessity. Russian military analysts have identified AI's military functions as ranging from auxiliary systems like logistics management, forces maintenance, and information security, to more central functions such as air defence and decision-making. In envisioning the future role of AI, some analysts anticipate wider integration of AI and suggest that the technology should be embedded into the entire command structure rather than merely addressing isolated tasks.[101] When assessing early warning and ISR capabilities, AI is seen as particularly valuable in specialised areas like radar information processing, where the risks of system malfunctions are comparatively small. However, some experts are cautious about its potential impact on crisis stability, even if AI is used in relatively limited areas. Meanwhile, discussions on command and control functions in open-source platforms remain very limited. Moreover, although there is ongoing discussion about the depth of AI's role in decision-making processes, there's a general agreement that humans should always play a central role in decisions on the use of nuclear weapons.

The **UK** views AI integration in military systems as both a strategic advantage and a responsibility in terms of using it ethically and legally. It acknowledges the potential of AI to bolster strategic stability, especially in decision-support tools and enhancing situational awareness through data analysis and information processes. Yet, it also recognises the risks the technology poses. Official sources highlight that the UK is concerned about the potential uses of AI by adversaries, especially when AI falls into the hands of adversaries who might use AI unethically, such as for social and population control. The UK is committed to using AI responsibly in a way that aligns with the country's democratic values, while also prioritising being at the forefront of AI development. Even though the country does not explicitly link AI with the nuclear weapons domain, it has emphasised that the human element will always be an integral part of critical decision-making processes. This underscores the nation's stance on maintaining political human control over nuclear weapons due to the fact that the use of AI in military operations raises accountability questions as well as risks associated with technological limitations. As such, London champions the establishment of clear norms of use and positive obligations as the cornerstone of ensuring human control over AI.

The **US** is also pursuing AI-enabled capabilities for operational and strategic superiority, in particular to maintain its military superiority over adversaries. The interest in further automating some NC3 applications relates to the need and willingness of augmenting leadership decision-making time, optimising early warning systems, speeding up order dissemination, and enhancing counterforce capabilities, which encompasses target identification and tracking.[102] The US underscores the necessity of harnessing this technology in ways that are lawful, ethical, responsible, and accountable.[103] The US champions human involvement in critical decisions related to nuclear weapons, advocating for a 'human-in-the-loop' approach. Concurrently, the US recognises potential AI vulnerabilities, drawing attention to the risks of accidental launches and escalations.[104]

Russian military analysts identified AI's military functions, ranging from auxiliary systems like logistics management, forces maintenance, and information security, to more central functions such as air defence and decision-making.

Areas of agreement

All the nuclear-weapons states recognise some similar risks and benefits from the integration of AI in military and defence systems, as well as their impact on decision-making. In this regard, all the examined states consider AI particularly valuable in ISR functions and for early warnings, decision support systems and resilient communication pathways. All the examined nuclear-weapons states acknowledge the risk that human operators might excessively trust inaccurate or poisoned data provided by AI. On this, China, France, Russia, the United Kingdom, and the United States would agree: Data insufficiency, corruption, and bias, and cyber manipulation and attack, can mislead and undermine decision-making.

Furthermore, each nuclear-weapons state is concerned about the potential for adversaries to integrate advanced AI models into their military systems first, which – they fear – could create a strategic imbalance. The willingness to gain a strategic edge, as well as the importance of keeping NC3 systems technically efficient has prompted a sense of urgency among nuclear-weapons states to stay ahead in technological developments.

Even though official documents might bypass explicit nuclear mentions, open-source records infer an agreement among all states that complete automation of nuclear decisions is unacceptable. A human agent should always retain prime decision-making authority. Conversely, the use of AI in supportive procedures, such as early warning or enhancing communication pathways and cyber defences, is considered valuable for decision-making processes. Yet, the scope and implications of AI integration into auxiliary systems (those not directly linked to command and control) is scarcely debated, especially its potential impact on functions critical to the NC3 architecture.

Overall, the analysis of P5 states' perspectives reveals that AI should complement, not take over, strategic command and control. All P5 countries perceive AI as an invaluable analytical tool for (1) decision-support, (2) maintaining reliable and secure communication channels, (3) enhancing situational awareness by swiftly processing vast data arrays, and offering comprehensive threat assessments.

Areas of disagreement

Conceptions of 'humans in the loop'

Although P5 states agree that a complete automation of nuclear decision-making is unacceptable, these nations differ in their approaches to and interpretations of the 'humans in the loop' concept, which refers to the necessity for human oversight or intervention in critical decision-making systems. This concept is particularly important in NC3, where leaving effective decision-making to machines raises profound concerns. All states value human judgment in nuclear decision-making, yet their reliance on and integration of AI in their NC3 structures varies.

The US, for example, is clear about ensuring human agency in nuclear decisions.[105] Regardless of AI and automation

Open-source records infer an agreement among all states that complete automation of nuclear decisions is unacceptable. A human agent should always retain prime decision-making authority.

advancements, the final decision should rest with the human operators, according the US perspective.[106] Likewise, the UK and France stress the need for human intervention in their nuclear decision-making, striving for a balance between technological advancements and human judgment.[107]

China's offical perspective of human-machine interaction remains somewhat ambiguous. Although China highlights the importance of maintaining human control over weapons systems, official records don't specify AI's role when it comes to nuclear weapons. However, unofficial discussions among scholars, policy experts, and some PLA officers hint at a prevailing belief that humans will remain central to strategic decision-making, with AI only remaining in a supportive role.

Russia's approach is even less transparent than China's. While Russia values the role of human judgment in the nuclear decision-making process, the precise nature and extent of human control in such systems remains subjects of debate.[108]

Nuclear postures and doctrines

The integration of AI into NC3 systems across P5 states is deeply influenced by diverse nuclear postures and doctrines.

For instance, a nuclear No First Use (NFU) policy means a commitment to using nuclear weapons only in response to a nuclear attack. Given the defensive nature underpinning a NFU doctrine, and its inherent focus on retaliation, the integration of AI in NC3 would likely be more skewed towards early warning systems, rapid response mechanisms, and precise delivery of nuclear assets for a potential counterattack. This would prioritise early and accurate detection of any incoming first strike. The specific situation with China, however, adds complexity to this narrative. Despite China's historical adherence to its NFU policy, the country's ongoing modernisation and nuclear build-up makes experts question the continuous viability of China's NFU policy.[109] Analysts and governments have warned that the country might be implementing a launch-on-warning posture, which contradicts their assumed practice of keeping warheads and delivery systems separate during peacetime.[110] Thus, while under the NFU policy, China's AI integration would largely revolve around retaliation and survivability of its second-strike capabilities, potential policy shifts might alter AI's role to accommodate new strategic priorities in the future.

Although P5 states agree that a complete automation of nuclear decision-making is unacceptable, these nations differ in their approaches to and interpretations of the 'humans in the loop' concept.

On the other hand, countries with a nuclear policy that does not rule out the pre-emptive use of nuclear weapons in a conflict under specific circumstances (as is the case with the US, Russia, UK and France), might prioritise AI-enhanced early warning systems to detect any potential threats quickly, providing the necessary window for a potential limited nuclear strike. Given the potential to deploy nuclear weapons early during conflicts, it is plausible that AI might be seen as valuable to ensure rapid decision support, which could include streamlined decision pathways to ensure rapid retaliation or escalation if needed.[111]

For instance, Russia's doctrine contemplates nuclear weapons use only in situations posing an existential threat to the nation. Russia's official stance is to resort to nuclear weapons only following

a nuclear or other weapons of mass destruction attack or if confronted with significant conventional aggression that threatens the country's existence. Some Western officials and analysts, however, believe that Russian leaders might deploy nuclear weapons early in a conflict in a limited manner, with the expectation of swiftly resolving the conflict in Russia's favour (tactic known as 'escalate to de-escalate').[112] Russia might thus see the value of AI-enhanced early warnings for rapid and accurate threat detection, as well as decision-support to possibly evaluate courses of action. Moreover, Russian nuclear doctrine views any attack on its NC3 as grounds for potentially deploying nuclear weapons. Thus, Russia is more concerned about ensuring its NC3 systems remain operational post-attack.[113] This ensures all regions, including remote areas, maintain communication and prevent disjointed reactions.

The US nuclear posture focuses on deterrence of nuclear and non-nuclear threats "with potential strategic effect for which nuclear weapons are necessary to deter"[114], assurance of allies and partners, and, if necessary, "the use of nuclear weapons in extreme circumstances to defend the vital interests of the United States or its Allies and partners".[115] The US nuclear triad emphasises strategic deterrence with capabilities for a full spectrum of nuclear conflict.

France's nuclear strategy is rooted in its commitment to minimal deterrence, also termed as 'strict sufficiency', wherein France keeps only those nuclear forces essential for deterring threats to its core interests. The country's nuclear capabilities are designed for the capacity to deliver unacceptably severe consequences to aggressors targeting France's critical interests. France maintains a clear stance: its nuclear forces are not aimed at any specific nation, and nuclear weapons are not seen as tools of war, but only as means to deter aggression.[116] However, France reserves the right to use nuclear weapons to protect what it terms as its vital interests. France, with its two-component (or dyad) nuclear force structure, might prioritise strategies focused more on assured retaliation. Key measures to achieve the resilience of French Armed Forces include the protection of AI-critical assets, resilient communication, self-reliant AI systems, and, in extreme situations, operational continuity without AI.[117]

The UK's deterrence strategy, the 'Continuous at-sea deterrence', relies on at least one ballistic missile submarine being on patrol at all times, ensuring a credible second-strike capability. Similar to France, the UK subscribes to the principle of maintaining the minimum credible nuclear deterrence, although in 2021 the country announced that, given the changing security environment, the cap on the nuclear stockpile will be raised.[118] Given the UK's nuclear posture, and similar to France, it is likely that AI can be utilised in decision support, enhance early-warning systems, and strengthen secure communication lines with patrolling submarines.[119]

Transparency

When it comes to publicly available information on the debate within the P5 states, most of the existing literature predominantly centres on the US. Historically, the US has upheld a degree of openness regarding its military and nuclear strategies compared to that of other nuclear states, particularly Russia and China. For

example, the US Department of Defense released its directive called "'Autonomy in Weapons Systems', in January 2023, which seeks to outline guidelines for ethical uses of AI in the military.[120]

The UK's Ministry of Defence has been relatively open about certain aspects of its defence planning. However, there's little to no mention of how AI relates to the UK's nuclear strategy, with the only caveat that the UK has committed to always maintain human control over nuclear weapons.[121]

Information about AI's role within France's nuclear systems is scarce. Only few independent analyses address the direct connection between AI and nuclear strategy in France.[122] Even with limited records, most insights come from independent experts. France has also committed to maintain human oversight over critical decision-making processes.[123]

China is known for its opacity regarding defence and security matters. Thus, a significant portion of understanding and insights about Chinese approaches to AI in NC3 comes from external analysts, indirect hints from official documents, or extrapolations from broader technological advancements.

Russia's discussions on AI and its nuclear dimension are often deeply embedded in its broader military debate. Although they discuss AI-based weaponry, details about NC3 integration are not disclosed. Analysts often have to interpret remarks from high-ranking officials, military doctrines, and defence industry to understand Russia's approach to AI in NC3.

Civil-military relations

The nature of civil-military relations varies widely among P5 states, which impacts directly how each state approaches and integrates AI into their NC3 systems.

Both the US and China emphasise the importance of collaboration between the public and private sectors for tech advancements, but in very different ways given the fundamentally different natures of their political systems. Given the significance of their computer industries in AI development, both countries aim to leverage them for defence and strategic gains. However, the 'Military-Civil Fusion' strategy by the Chinese Communist Party aims to merge civilian tech innovations with military applications. This strategy enables the Chinese Party to reassess the Chinese science and technology sector to ensure that technological innovations benefit both the economic and military domains. Unlike the Western approach which distinguishes civilian tech from military use, China promotes a significantly more integrated approach.

Both the US and China emphasise the importance of collaboration between the public and private sectors for tech advancements, but in very different ways given the fundamentally different natures of their political systems.

Russia's approach to technology, especially in domains concerning national security, is highly state-centric. The Kremlin maintains tight control over military developments. Russia's state-driven model of development means that the military and strategic sectors can continue to receive funding, especially for projects deemed critical to national security even as the commercial AI sector faces challenges, partly due to the post-Cold War economic struggles, but also due to the Russian war of aggression against Ukraine and subsequent Western sanctions that caused talent drain and reduced investments.[124]

Ethical considerations

Ethical considerations surrounding the use of AI in weapons systems are complex and cross-cutting, touching upon areas such as decision-making autonomy, control over lethal force, and the potential consequences of misjudgements. The US, UK, and France have been more publicly vocal than Russia and China about their ethical stances. For instance, the Defense Innovation Board, an advisory group to the US Department of Defense, has provided recommendations on the ethical use of AI in military operations, emphasising human control and oversight.[125] The UK Ministry of Defence has articulated principles for AI's ethical use in military settings and has pledged to use AI to uphold democratic values, in contrast with the use of AI from some states for social and population control.[126] France has also taken steps at the national level to address AI's ethical challenges.[127]

Accusations are rampant, with nations claiming possibly unethical use of AI by adversaries. For instance, in its 'Defence Artificial Intelligence Strategy', the UK claimed that;

> "Adversaries and systemic competitors are investing heavily in AI technologies to challenge our defence and security edge. We have already seen claims that current conflicts have been used as test beds for AI-enabled autonomous systems, and we know that adversaries will use technology in ways that we would consider unethical and unsafe. Potential threats include enhanced Cyber and information warfare, AI-enabled surveillance and population control, accelerated military operations and the use of autonomous physical systems."[128]

Russia acknowledges the benefits of AI in military systems and there is significantly less emphasis on ethical considerations in the Russian literature compared to Western nuclear-weapons states. Officially, China emphasises the peaceful use of AI technologies.[129] However, its rapid military modernisation, including the integration of AI into various military platforms, suggests a pragmatic approach. Ethical considerations are discussed, both in official statements and in academic circles, but how they specifically apply to military decision-making systems is less clear from official documents.[130]

Recommendations

Given the challenges and opportunities highlighted at the ELN workshops on the intersection of AI and nuclear decision-making, it is imperative for the P5 states to take proactive steps to address risks posed by AI in nuclear decision-making. These recommendations identified during the ELN project can help to mitigate the risks of integrating AI into NC3 systems.

Multilateral approaches:

- P5 states should be transparent about their AI risk management plans and detail them in national reports, for example to NPT meetings of states parties. This would not only bolster trust but also highlight how they tackle risks associated with AI in the nuclear field, particularly decision-making.

- NPT states parties should agree to host multistakeholder sessions in this NPT review cycle that delve into AI developments. By incorporating private sector and academia perspectives, they could ensure a comprehensive view of the risks and possible solutions at hand. On 26 October 2023, the UN Secretary General António Guterres announced the creation of a new AI Advisory Body with the goal to create preliminary recommendations on international governance of AI and a shared understanding of risks posed by this technology, among others.[131] This scientific advisory body could be instrumental in informing States Parties about the potential impacts of AI in NC3 during the eleventh NPT review cycle.

- Building on the common ground and reflected in a paragraph of the draft final document of the 2020 NPT Review Conference, the nuclear-weapons states should strengthen their commitment to foster dialogues amongst themselves and with non-nuclear states to delve into the impact of emerging technologies on the nuclear landscape. Such efforts should involve a 'fear mapping' to identify and communicate their concerns over nuclear AI advancements.[132] UNODA or UNIDIR could take on the task of designing and collating such reports. The 'Creating an Environment for Nuclear Disarmament' (CEND) initiative could facilitate discussions surrounding AI risks.

- The nuclear-weapon states should prioritise regular dialogues on AI's influence over nuclear decision-making to reduce risks generated by AI's potential to disrupt stability. Sharing methodologies, especially concerning reliability safeguards, could further strengthen confidence among the nuclear-weapon states.[133]

- P5 states should consider establishing standardised red-teaming methodologies.[134] In particular, a centralised red-teaming lab where countries can collectively evaluate and test their AI applications against vulnerabilities should be created.

I. Profiling AI risks within NC3:

P5 states should develop a comprehensive analytical framework to evaluate the risks of integration of AI risks within each NC3 application. They can do so by building on the initial framework provided below.

P5 states should develop a comprehensive analytical framework to evaluate the risks of integration of AI risks within each NC3 application. They can do so by building on the initial framework presented in this report.

The following framework can be used by P5 states to establish thresholds to prevent high-risks AI integrations. This aligns with President's Biden 'Executive order on safe, secure and trustworthy artificial intelligence', which establishes new standards for AI safety and security.[135]

The following framework provides an initial concept for how metrics can be developed based on parameters that impact models' risks of integration. It is applicable to current advanced deep learning models (such as generative AI and LLMs); however, it is also designed to be 'model agnostic' and work with future models even if a different paradigm emerges presenting similar technological risks. Moreover, by changing the integration risks, this framework can be modified to work with other emerging risks.

The metrics are divided into two overarching areas: (1) technological risks, which encompass the technology's capabilities, transparency, and reliability, and (2) integration risks, focusing on the area of integration and automation degree. The first metrics consider various technological aspects, such as a model's adaptability across different tasks, domains, and methods; the model's scalability based on architecture, which examines how effectively it can grow with more data and computational power; and the overall performance of the model. There's a clear link between these aspects and the risks associated with a model. As models scale and perform better, they introduce unique risks. These risks are closely tied to measures of transparency and reliability, which account for issues like erroneous AI outputs (known as 'hallucinations'), cyber threats, and alignment risks.[136]

Transparency relates to human's understanding of the algorithms a model uses to produce an output. Models that are less transparent, but have a wide scope and superior performance, present different risks compared to those that are more transparent and have a narrow range of tasks.

Reliability risks are associated with behaviors of the model that provide direct risks. These include hallucinations, which are incorrect outputs generated by the model with a high degree of confidence; cyber security risks, which relate to the ability of malicious actors to modify the behavior of the model through manipulating data, vulnerabilities, and attacks; alignment, which relates to the risks associated with aligning a model to human preferences and our ability to control and steer a model's outputs. Reliability risks are heavily related to performance and transparency. Risks associated with reliability grow in proportion to the technological risk.

Furthermore, integration risks evaluate the consequences of incorporating the model in a specific NC3 area and the extent of the automation required for the model to perform.

AI technological risks encompass:

- Transparency (the interpretability of the architecture or model degree to which an AI model arrives to a certain conclusion can be understood).

- Generalisability (the AI's adaptability, referring to how versatile the AI is when faced with different tasks, domains, or methods).

- Performance (the AI's competency across tasks and domains determined by benchmarking. In other words; how well the AI does the job).
- Scalability (the scaling laws of the architecture determined by a combination of parameters, training data, and compute. In other words; a model's ability to improve or maintain its performance as it encounters more data or is given additional computational resources).[137]
- Reliability. Factors include:
    - Susceptibility to generating erroneous outputs with unwavering confidence, or 'hallucinations'.
    - Susceptibility to cyber security vulnerabilities.
    - Alignment risks (the degree to which the AI is aligned with human values and the control that we have over the systems behaviour).

NC3 Integration risks are divided as follows:

- Area of integration in NC3:
    - Low impact, such as monitoring system health, communication path optimisation, and training.
    - Moderate impact, such as sensor data analysis, enhancing cyber security defenses, logistics.
    - High impact, such as suggesting courses of action during a crisis, identifying prelaunch activities, target identification.
    - Very high impact, such as autonomous retaliation.
- Level of system's autonomy (the degree of human supervision required in a specific NC3 area):
    - Light automation: automation of one-related task.
    - Semi-autonomous: systems with partial automation, but crucial decisions require human intervention.
    - Full autonomy: systems that can make decisions and operate independently without human intervention.

These metrics can help P5 states to gauge the risk levels of different AI systems. Presently, the most concerning risks arise from cutting-edge deep learning models, particularly those based on the transformer architecture like LLMs and other foundational models. As these models evolve, their reliability risks also increase, especially when used in high-impact decision-making situations and with a significant degree of automation. While these models are impressively versatile and adaptable across domains without performance degradation, they also exhibit a worrying unreliability with significant rate of erroneous outputs and their 'black box' nature makes it difficult to understand why they arrive at certain conclusions. As their capabilities expand, ensuring they align with

human values becomes crucial. Misalignments, when integrated into critical systems, can lead to grave errors.

Conversely, more specialised systems can also pose risks if they lack reliability and integration risks are high. For example, machine vision models that have low transparency and high risk of hallucination provide high levels of risk if integrated in systems related to satellite image analysis. The risks are heavily dependent on the area of integration. This demonstrates how the integration area in NC3 significantly influences these risks.

II. Advancing a moratorium on high-risks models in NC3 systems

Using the risk assessment framework produced in the prior chapter, and considering the significant risks posed by cutting-edge deep learning models, particularly LLMs and other foundational models, P5 states should impose a moratorium on the integration of these models into specific NC3 areas with the highest integration risks.

The risk profiling system introduced here can be elaborated to provide a scoring system to enable the classification of high-risk AI systems. This can be used as the basis for a moratorium. Understanding the risks associated with these models, makes it evident that there's a pressing need to enforce a moratorium on integrating them into NC3 systems.

The moratorium should remain in place until such technologies become entirely interpretable and until the technological risks with these models is solved. Instituting a moratorium is not just about pausing; it's a strategic move to mitigate the embedded risks and uncertainties of such integration.[138]

By advocating for this moratorium, the P5 states can create a buffer to assess the technology's strengths, weaknesses, and implications for global security. Beyond that, such a pause will act as a catalyst for global dialogue, generating discussions on aspects such as ethical standards, rules of engagement, and methods of control and verification for AI-integrated NC3 systems.

Looking ahead, UN Secretary-General Antonio Guterres has identified the upcoming Summit for the Future, scheduled for September 2024, as the suggested forum for deliberation on the impact of AI on peace and security. In this context, the UN Secretary-General also mentioned potential nuclear implications. Building on the foundation of the SDG Summit, the Summit for Future aims to anchor its vision in the ideals of the United Nations Charter and the 2030 Agenda. Given the summit's forward-looking theme, it stands a relevant platform to broach the topic of the moratorium.

Furthermore, the REAIM summit, set to convene in Seoul late in 2024, emerges as another significant platform for discussion. With its diverse participation encompassing officials, non-governmental agencies, academia, and the private sector, the summit involves diverse insights that can offer holistic strategies to address AI and its nuclear nexus. This collaboration with non-governmental actors finds resonance in the UN's recent initiatives. Specifically, Secretary-General Guterres' announcement of a High-Level Advisory Board for Artificial Intelligence showcases the importance of a multi-stakeholder approach. The board's forthcoming insights

Considering the significant risks posed by cutting-edge deep learning models, P5 states should impose a moratorium on the integration of these models into specific NC3 areas with the highest integration risks.

on pathways for global AI governance will undoubtedly serve as a reference point in navigating the challenges ahead.

Undoubtedly, the private sector holds a critical role in shaping the discussion. Their firsthand experiences and depth of knowledge can provide practical viewpoints for potential policies' implementation. Moving forward, it's imperative to combine diverse viewpoints to craft a cohesive strategy.

Finally, for AI models that do not carry the same high-level risks as state-of-the-art deep-learning systems, it's crucial for P5 states to maintain a unified position on retaining human control in nuclear decision-making. Such a shared commitment to keep a 'human in the loop' during any decision to launch a nuclear weapon could build on the January 2022 commitment by the P5 to "consider the avoidance of war between Nuclear-Weapon States and the reduction of strategic risks as [their] foremost responsibilities."[139]

Bilateral initiatives:

The US-China and US-Russia bilateral relationships stand as paramount channels for mitigating AI-related risks. While a top-down approach is indispensable, it's equally crucial to create awareness and common ground on the risk associated with the integration of AI into NC3 from the bottom-up:

- For AI models that do not carry the same high-level risks, bilateral intergovernmental discussions at the track-1 level should revolve around the retention of human control over nuclear systems and the peril of becoming overly reliant on automation.

- In parallel, track-2 dialogue could delve into technical subjects, such as practical ways to assure human oversight in real-world AI applications and devising rigorous testing procedures that can validate the reliability and predictability of AI systems in nuclear contexts.[140] The private sector, with its cutting-edge knowledge of AI's risks, should be involved in informing these discussions.

The US-China and US-Russia bilateral relationships stand as paramount channels for mitigating AI-related risks. While a top-down approach is indispensable, it's equally crucial to create awareness and common ground on the risk associated with the integration of AI into NC3 from the bottom-up.

Conclusions

The study conducted by the ELN underscores the imperative for continued research to develop a robust and holistic risk profiling framework. While this research report laid the foundation through an initial framework, delving deeper into the intricacies of cutting-edge models is paramount. This gap presents a promising avenue for subsequent studies.

Moreover, this study highlights the importance of encompassing insights from a spectrum of stakeholders. Industry experts and academic scholars, with their specialised knowledge, bring invaluable perspectives in informing these discussions. Ensuring their involvement will be crucial in crafting comprehensive and effective risk profiles.

Annex: Key terms glossary

Term	Definition
AI	Artificial Intelligence (AI) is the simulation of human intelligence processes by machines. The goals of the field include reasoning, knowledge representation, planning, learning, natural language processing, perception, robotics, and achieving general intelligence. The ultimate aspiration of AI is to build systems that can perform all tasks that would require human intelligence.[141]
Deep Learning	Deep learning networks have multiple (deep) layers between the input and the output, allowing for more complex patterns and representations to be learned. The depth of these networks allows for the extraction of hierarchically structured features. For example, in image processing tasks, the initial layers might learn to detect edges, the middle layers might learn to detect shapes by combining edges, and the deeper layers might learn to detect more complex structures.[142]
Foundation model	Foundation models, sometimes referred to as base models, are large models trained on massive datasets, usually via self-supervised or semi-supervised learning. The strength of foundation models is in their ability to be further tuned and adapted to specialised tasks.[143]
Generative AI	Generative artificial intelligence is a subset of artificial intelligence designed to produce new content including text, images, or other forms of media. These deep learning models are trained on large datasets, which allow them to discern patterns, styles, and structures inherent in the data. Once trained, they can create novel outputs that align with the characteristics of the training data.[144]
Large Language Model	Large Language Models (LLMs) are deep learning models that excel in understanding and generating human language. Their sheer size, encompassing billions of parameters, gives them the capacity to process language with general-purpose flexibility. Most LLMs are constructed using the transformer architecture, predominantly the autoregressive models which predict subsequent tokens based on given input.[145]
Machine Learning	Machine Learning (ML) is a subset of artificial intelligence that enables systems to improve and adapt from experience without being explicitly programmed. It is characterised by the machine's ability to autonomously derive algorithms from data, bypassing the traditional need for human-crafted algorithms. The primary learning paradigms for machine learning are supervised, semi-supervised, and unsupervised learning.[146]
Neural Networks	An artificial neural network (ANN) is a computational model inspired by the way biological neural networks in the human brain function. At its simplest form, an ANN comprises nodes or 'neurons' which process information and are interconnected through 'synapses' with certain weights. These weights are adjusted during training to minimise the difference between the predicted output and the actual target values. The foundational building block of an ANN is the neuron or node. Each neuron takes multiple inputs, processes them, and produces an output. The process typically involves taking a weighted sum of the inputs, adding a bias, and then passing it through a non-linear activation function.[147]

Semi-supervised Learning	Semi-supervised learning is positioned between supervised and unsupervised learning methodologies. It uses both labeled and unlabeled data for training. The idea is to leverage the large amount of available unlabeled data to enhance the learning performance obtained with the smaller set of labeled data. This approach can be particularly advantageous when acquiring labeled data is expensive or labor-intensive, but there's ample unlabeled data. The model benefits from the labeled data for precise guidance, while also capitalising on the insights derived from the broader scope provided by the unlabeled data.[148]
Supervised Learning	Supervised learning is a machine learning paradigm where the model is trained on a labeled dataset. This means that for every input in the training dataset, the correct output is known. The goal of supervised learning is to learn a mapping from inputs to outputs and make predictions on new, unseen data. This approach is analogous to learning with a teacher, where the teacher provides guidance and the correct answers, and the model iteratively adjusts to approximate this mapping. Typical applications include regression (predicting a continuous value) and classification (categorising data into predefined classes).[149]
Transformer	The transformer architecture is a deep learning model introduced in 2017. It leverages attention mechanisms to process input data in parallel, rather than in sequence, allowing it to efficiently handle long-range dependencies and achieve state-of-the-art performance in tasks like machine translation and text summarisation.[150] All transformer models have: 1. **Tokenisers:** Convert text into smaller chunks called tokens. 2. **Embedding Layer:** Maps tokens and their positions to vectors, capturing their meaning in a continuous space. 3. **Transformer Layers:** Sequentially refines these vectors, diving deeper into linguistic nuances. Each layer contains: • **Attention mechanisms:** Which weigh the importance of different tokens relative to each other. • **Feedforward layers:** Which further transform the data.
Unsupervised Learning	Unsupervised learning deals with datasets that lack labeled responses. Without a teacher or guide, the model is tasked with finding the inherent structure or patterns within the data. Essentially, the algorithm teaches itself by analysing the data's underlying structure and relationships. Common techniques include clustering, where data is grouped based on similarities, and dimensionality reduction, which condenses information into fewer variables while retaining most of its variance. Unsupervised learning is pivotal when the nature of the data is unknown, or when labeling data is costly or time-consuming.[151]

References

1 The term 'interpretability' refers to the ability to understand and explain how a machine learning model or AI system arrives at a particular decision or output. It involves making the inner workings, reasoning, and decision-making processes of the AI system transparent and comprehensible to humans. For more information see Tim G.J Rudner & Helen Toner, "Key Concepts in AI safety: Interpretability in Machine Learning", *Center for Security and Emerging Technology* (March 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-interpretability-in-machine-learning.

2 Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, & Shirui Pan, "Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning", *arXiv* (October 2023), http://arxiv.org/abs/2310.01061.

3 Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", SIPRI (June 2020), p. ix, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_str ategic_stability_and_nuclear_risk.pdf.

4 Samuel Bendett, "Russian Military Debates AI Development and Use", The Azure Forum (4 May 2023), https://www.azureforum.org/russian-military-debates-ai-development-and-use/.

5 *Military and Security Developments Involving the People's Republic of China* (Washington, D.C.: Department of Defense, 2022), p. v, https://media.defense.gov/ 2022/Nov/29/2003122279/-1/-1/1/2022-MILITARY-AND-SECURITY-DEVELO PMENTS-INVOLVING-THE-PEOPLES-REPUBLIC-OF-CHINA.PDF.

6 *Defence Artificial Intelligence Strategy* (London: UK Ministry of Defence, June 2022), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf ; *Artificial Intelligence in Support of Defence: Report of the AI Task Force, Ministère des Armées* (Paris: French Ministry of the Armed Forces, September 2019), https://www.defense.gouv.fr/sites/default/files/aid/Report%20of%20the%20AI%20Task%20Force%20September%202019.pdf.

7 Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", SIPRI (June 2020), p. 64, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_str ategic_stability_and_nuclear_risk.pdf.

8 The term "inadvertent escalation" describes situations where actions, though intentional, lead to unintended escalation. "Accidental escalation" pertains to outcomes arising from events that were never intended to escalate. For more information see Forrest E. Morgan, Karl P. Mueller, Evan S. Medeiros, Kevin L. Pollpeter, & Roger Cliff, "The Nature of Escalation", in: Forrest E. Morgan et al. (eds.), *Dangerous Thresholds: Managing Escalation in the 21st Century*, 1st ed. (Santa Monica: RAND Corporation, 2008), pp. 7-46, http://www.jstor.org/stable/10.7249/mg614af.9.

9 See Rafael Loss & Joseph Johnson, "Will Artificial Intelligence Imperil Nuclear Deterrence", War on the Rocks (19 September 2019), https://warontherocks.com/2019/09/will-artificial-intelligence-imperil-nuclear-deterrence/ ; James Johnson, 'Artificial Intelligence and Nuclear Risk: Rethinking Deterrence Strategy in the Age of AI' in: *Understanding the humanitarian consequences and risks of nuclear weapons* (Vienna: Federal Ministry Republic of Austria, July 2023), pp. 21-30 (p. 22).

10 Wyatt Hoffman and Heeu Millie Kim, "Reducing the Risks of Artificial Intelligence for Military Decision Advantage", *Center for Security and Emerging Technology* (March 2023), p. 4, https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/.

11 Izumi Nakamitsu, "Opening remarks by Ms. Izumi Nakamitsu High Representative for Disarmament Affairs at the First Preparatory Committee for the 2026 Review Conference of the Treaty on the Non-Proliferation of Nuclear Weapons" (Vienna: 31 July 2023).

12 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures", *Institute for Security and Technology* (February 2023), https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf ; Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", *SIPRI* (June 2020), https://www.sipri.org/sites/default/ files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf ; Jill Hruby & M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems", *NTI Paper* (August 2021), https://www.nti.org/wp-content/uploads/2021/09/NTI_Paper_AI_r4.pdf ; James Johnson, *AI and the Bomb: Nuclear strategy and risk in the digital age* (Oxford: Oxford University Press, 2023) ; Tim McDonnell, Mary Chesnut, Tim Ditter, Anya Fink & Larry Lewis, "Artificial Intelligence in Nuclear Operations: Challenges, opportunities, and impacts", *Center for Naval Analyses* (April 2023), https://www.cna.org/reports/2023/04/Artificial-Intelligence-in-Nuclear-Operations.pdf ; Michael T. Klare, "Skynet Revisited: The Dangerous Allure of Nuclear Command Automation", *Arms Control Today*, (April 2020) https://www.armscontrol.org/act/2020-04/features/skynet-revisited-dangerous-allure-nuclear-command-automation.

13 Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", *SIPRI* (June 2020), p.19, https://www.sipri.org/sites/default/ files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf

14 Jill Hruby & M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems", *NTI Paper* (August 2021), p.6, https://www.nti.org/wp-content/uploads/2021/09/NTI_Paper_AI_r4.pdf ; Michael Horowitz, Paul Scharre, & Alexander Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence", *arXiv* (December 2019), p.6, arxiv.org/abs/1912.05291.

15 "Machine Learning Paradigms", in: *Introduction to Machine Learning*, (Champaign: Wolfram Research), https://www.wolfram.com/language/introduction-machine-learning/machine-learning-paradigms/.

16 "Computational approaches to Explainable Artificial Intelligence: Advances in theory, applications and trends", Information fusion (2023), https://www.sciencedirect.com/science/article/pii/S1566253523002610#:~:text=Deep%20Learning%20(DL)%2C%20a,high%2Dlevel%20features%20from%20data.

17 Helen Toner, "What Are Generative AI, Large Language Models, and Foundation Models?", *Center for Security and Emerging Technology* (12 May 2023), https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/.

18 Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, & Ahmed Awadallah, "Orca: Progressive Learning from Complex Explanation Traces of GPT-4", *arXiv* (June 2023), https://arxiv.org/abs/2306.02707.

19 'Parallelism', in the context of attention mechanisms, refers to processing multiple pieces of data concurrently rather than sequentially. This greatly speeds up the processing time.

20 Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, & Furu Wei, "Retentive Network: A Successor to Transformer for Large Language Models", *arXiv* (August 2023), https://arxiv.org/abs/2307.08621.

21 "Explainable Artificial Intelligence (XAI)", *Defense Advanced Research Projects Agency*, https://www.darpa.mil/program/explainable-artificial-intelligence

22 Melanie Mitchell & David C. Krakauer, "The debate over understanding in AI's large language models", *Proceedings of the National Academy of Sciences of the United States of America* (28 March 2023), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10068812/

23 See Rafael Loss & Joseph Johnson, "Will Artificial Intelligence Imperil Nuclear Deterrence", *War on the Rocks* (19 September 2019), https://warontherocks.com/2019/09/will-artificial-intelligence-imperil-nuclear-deterrence/ ; James Johnson, 'Artificial Intelligence and Nuclear Risk: Rethinking Deterrence Strategy in the Age of AI' in: *Understanding the humanitarian consequences and risks of nuclear weapons* (Vienna: Federal Ministry Republic of Austria, July 2023), pp. 21-30 (p. 22).

24 See Philip Reiner, Alexa Wehsener, & M. Nina Miller, "When Machine Learning Comes to Nuclear Communication Systems", *C4ISRNET*, (1 May 2020) https://www.c4isrnet.com/thought-leadership/2020/04/30/when-machine-learning-comes-to-nuclear-communication-systems/ ; Jessica Cox & Heather Williams, "The Unavoidable Technology: How Artificial Intelligence Can Strengthen Nuclear Stability", *The Washington Quarterly*, 44:1 (2021) ; Philip Reiner & Alexa Wehsener, "The real value of artificial intelligence in nuclear command and control', *War on the rocks* (4 November 2019), https://warontherocks.com/2019/11/the-real-value-of-artificial-intelligence-in-nuclear-command-and-control/.

25 See Michael Klare & Chris Rostampour, "ChatGPT Sparks US Debate Over Military Use of AI", *Arms Control Association*, (June 2023) https://www.armscontrol.org/act/2023-06/news/chatgpt-sparks-us-debate-over-military-use-ai ; Josh Baughman, "China's ChatGPT War", *China Aerospace Studies Institute* (21 August 2023), https://www.airuniversity.af.edu/Portals/10/CASI/documents/Research/Cyber/2023-08-21%20China's%20ChatGPT%20War.pdf.

26 Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", *SIPRI* (June 2020), p.19, https://www.sipri.org/sites/default/ files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

27 Ibid.

28 Bruce G. Blair, "Loose cannons: The president and US nuclear posture", *Bulletin of the Atomic Scientists*, vol. 76, no. 1 (Jan. 2020), pp. 14–26, p. 15 ; Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", *SIPRI* (June 2020), p.19, https://www.sipri.org/sites/default/ files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

29 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures", *Institute for Security and Technology* (February 2023), p. 9, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

30 Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", *SIPRI* (June 2020), p.20, https://www.sipri.org/sites/default/ files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

31 Ibid, p. 21.

32 Jill Hruby & M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems", *NTI Paper* (August 2021), p. 15, https://www.nti.org/wp-content/uploads/2021/09/NTI_Paper_AI_r4.pdf.

33 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence

Building Measures", *Institute for Security and Technology* (February 2023), p. 9, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

34 Jill Hruby & M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems", *NTI Paper* (August 2021), p. 12, https://www.nti.org/wp-content/uploads/2021/09/NTI_Paper_AI_r4.pdf.

35 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures", *Institute for Security and Technology* (February 2023), p. 10, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

36 See Jill Hruby & M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems", *NTI Paper* (August 2021), https://www.nti.org/wp-content/uploads/2021/09/NTI_Paper_AI_r4.pdf ; Peter Rautenbach, "Artificial Intelligence and Nuclear Command, Control, and Communications: The Risks of Integration", *Effective Altruism Forum* (18 November 2022), https://forum.effectivealtruism.org/posts/BGFk3fZF36i7kpwWM/artificial-intelligence-and-nuclear-command-control-and-1#4_2_1_Automation_Bias.

37 James Johnson, "Automating the OODA loop in the age of intelligent machines: reaffirming the role of humans in command-and-control decision-making in the digital age", *Defence Studies*, 23:1, 43-67, p. 53, https://www.tandfonline.com/doi/full/10.1080/14702436.2022.2102486.

38 Jill Hruby & M. Nina Miller, "Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems", *NTI Paper* (August 2021), p. 21, https://www.nti.org/wp-content/uploads/2021/09/NTI_Paper_AI_r4.pdf.

39 Peter Rautenbach, "Artificial Intelligence and Nuclear Command, Control, and Communications: The Risks of Integration", *Effective Altruism Forum* (18 November 2022), https://forum.effectivealtruism.org/posts/BGFk3fZF36i7kpwWM/artificial-intelligence-and-nuclear-command-control-and-1#3_4_Predictive_Forecasting_of_the_Imminent_Use_of_Nuclear_Weapons.

40 James Johnson, "Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?", *Journal of Strategic Studies* 45(3):439–477, p. 441, https://doras.dcu.ie/25508/1/JSS%20JamesJohnson%20%282020%29.pdf.

41 Karen Weise & Cade Metz, "When A.I. Chatbots Hallucinate", *The New York Times* (1 May 2023), https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html.

42 Alice Saltini, "To Avoid Nuclear Instability, a Moratorium on Integrating AI into Nuclear Decision-making is Urgently Needed. The NPT PrepCom Can Serve as a Springboard", *European Leadership Network* (28 July 2023), https://www.europeanleadershipnetwork.org/commentary/to-avoid-nuclear-instability-a-moratorium-on-integrating-ai-into-nuclear-decision-making-is-urgently-needed-the-npt-prepcom-can-serve-as-a-springboard/.

43 Yuna Huh Wong, John M. Yurchak, Robert W. Button, Aaron Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris & Sebastian Joon Bae, "Deterrence in the Age of Thinking Machines", *RAND Corporation* (2020) p. xi, https://www.rand.org/pubs/research_reports/RR2797.html ; Peter Rautenbach, "Artificial Intelligence and Nuclear Command, Control, and Communications: The Risks of Integration", Effective Altruism Forum (18 November 2022), https://forum.effectivealtruism.org/posts/BGFk3fZF36i7kpwWM/artificial-intelligence-and-nuclear-command-control-and-1#4_2_3_Out_of_the_loop.

44 Yuna Huh Wong, John M. Yurchak, Robert W. Button, Aaron Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris & Sebastian Joon Bae, "Deterrence in the Age of Thinking Machines", *RAND Corporation* (2020) p. xi, https://www.rand.org/pubs/research_reports/RR2797.html.

45 Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", *SIPRI* (June 2020), pp. 104-105, https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

46 Maximilian Hoell & Sylvia Mishra, "Artificial Intelligence in Nuclear Command, Control, and Communications: Implications for the Nuclear Non-Proliferation Treaty" in: *The Implications of Emerging Technologies in the Euro-Atlantic Space: Views from the Younger Generation Leaders Network*, ed. by Julia Berghofer, Andrew Futter, Clemens Häusler, Maximilian Hoell & Juraj Nosál (Cham: Palgrave Macmillan, 2023), pp. 123 – 142 (pp. 135-136).

47 James Johnson, "Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?", *Journal of Strategic Studies* 45(3):439–477, p. 459, https://doras.dcu.ie/25508/1/JSS%20JamesJohnson%20%282020%29.pdf.

48 Wyatt Hoffman and Heeu Millie Kim, "Reducing the Risks of Artificial Intelligence for Military Decision Advantage", *Center for Security and Emerging Technology* (March 2023), p. 4, https://cset.georgetown.edu/publication/reducing-the-risks-of-artificial-intelligence-for-military-decision-advantage/.

49 Lei Gao & Ling Guan, "Interpretability of Machine Learning: Recent Advances and Future Prospects", *arXiv* (30 April 2023), https://arxiv.org/pdf/2305.00537.pdf.

50 Vincent Boulanin, Lora Saalman, Petr Topychkanov, Fei Su, & Moa Peldán Carlsson, "Artificial Intelligence, Strategic Stability and Nuclear Risk", *SIPRI* (June 2020), p. 107, https://www.sipri.org/sites/default/ files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf.

51 Michael Horowitz, "Artificial Intelligence and Nuclear Stability", in: Vincent Boulanin (ed.), The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk, volume I (Stockholm: SIPRI, May 2019), p. 82, https://www.sipri.org/publicati ons/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclearrisk-volume-i-euro-atlantic.

52 Peter Rautenbach, "Artificial Intelligence and Nuclear Command, Control, and Communications: The Risks of Integration", *Effective Altruism Forum* (18 November 2022), https://forum.effectivealtruism.org/posts/BGFk3fZF36i7kpwWM/artificial-intelligence-and-nuclear-command-control-and-1#2_1_Artificial_Intelligence_and_Machine_Learning.

53 *2022 Nuclear Posture Review* (Washington D.C.: U.S. Department of Defense, 2022), p. 22, https://fas.org/wp-content/uploads/2023/07/2022-Nuclear-Posture-Review.pdf.

54 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures", *Institute for Security and Technology* (February 2023), p. 9, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf ; Tim McDonnell, Mary Chesnut, Tim Ditter, Anya Fink & Larry Lewis, "Artificial Intelligence in Nuclear Operations: Challenges, opportunities, and impacts", *Center for Naval Analyses* (April 2023), p.10, https://www.cna.org/reports/2023/04/Artificial-Intelligence-in-Nuclear-Operations.pdf.

55 Alice Saltini, "British thinking of AI integration into and interaction with nuclear command and control, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

56 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

57 Ibid.

58 The State Council Information Office of the People's Republic of China, '中国的军事战略 [China Military Strategy]', (27 May 2015), https://english.www.gov.cn/archive/white_paper/2015/05/27/content_281475115610833.htm>.

59 Ibid.

60 Oleg Shakirov, "Russian thinking of AI integration into and interaction with nuclear C2, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

61 Oleg Shakirov, "Russian thinking of AI integration into and interaction with nuclear C2, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

62 *Summary of the 2018 Department of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity* (Washington, D.C.: Department of Defense, 2019), p. 7, https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF.

63 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures", *Institute for Security and Technology* (February 2023), p. 10, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf ; Tim McDonnell, Mary Chesnut, Tim Ditter, Anya Fink & Larry Lewis, "Artificial Intelligence in Nuclear Operations: Challenges, opportunities, and impacts", *Center for Naval Analyses* (April 2023), p.10, https://www.cna.org/reports/2023/04/Artificial-Intelligence-in-Nuclear-Operations.pdf.

64 *Defence Artificial Intelligence Strategy* (London: UK Ministry of Defence, June 2022), p. 32, https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy.

65 Alice Saltini, "British thinking of AI integration into and interaction with nuclear command and control, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

66 *Artificial Intelligence in Support of Defence: Report of the AI Task Force* (Paris: French Ministry of the Armed Forces, September 2019), p. 17, https://www.defense.gouv.fr/sites/default/files/aid/Report%20of%20the%20AI%20Task%20Force%20September%202019.pdf.

67 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

68 Fei Su & Jingdong Juan, "Chinese thinking on AI integration and interaction with nuclear C2, force structure, and decisionmaking", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

69 Oleg Shakirov, "Russian thinking of AI integration into and interaction with nuclear C2, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

70 *2022 Nuclear Posture Review* (Washington D.C.: U.S. Department of Defense, 2022), p. 22, https://fas.org/wp-content/uploads/2023/07/2022-Nuclear-Posture-Review.pdf.

71 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence

Building Measures", *Institute for Security and Technology* (February 2023), p. 11, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

72 *Defence Artificial Intelligence Strategy* (London: UK Ministry of Defence, June 2022), p. 10, https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy.

73 Ibid.

74 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", E*uropean Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

75 *Artificial Intelligence in Support of Defence: Report of the AI Task Force* (Paris: French Ministry of the Armed Forces, September 2019), p. 17, https://www.defense.gouv.fr/sites/default/files/aid/Report%20of%20the%20AI%20Task%20Force%20September%202019.pdf.

76 *The State Council Information Office of the People's Republic of China*, '中国的军事战略 [China Military Strategy]', (27 May 2015), https://english.www.gov.cn/archive/white_paper/2015/05/27/content_281475115610833.htm

77 Fei Su & Jingdong Juan, "Chinese thinking on AI integration and interaction with nuclear C2, force structure, and decisionmaking", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

78 Oleg Shakirov, "Russian thinking of AI integration into and interaction with nuclear C2, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

79 Based on testimony from Lt. Gen. Michael E. Kurilla at the Senate Armed Services Committee nomination hearing on 7 February 2022. See David Vergun, "Artificial Intelligence, Autonomy Will Play Crucial Role in Warfare, General Says", *U.S. Department of Defense* (8 February 2022,) https://www.defense.gov/News/News-Stories/Article/Article/2928194/artificial-intelligence-autonomy-will-play-crucial-role-in-warfare-general-says/.

80 Tim McDonnell, Mary Chesnut, Tim Ditter, Anya Fink & Larry Lewis, "Artificial Intelligence in Nuclear Operations: Challenges, opportunities, and impacts", *Center for Naval Analyses* (April 2023), p.10, https://www.cna.org/reports/2023/04/Artificial-Intelligence-in-Nuclear-Operations.pdf.

81 *Defence Artificial Intelligence Strategy* (London: UK Ministry of Defence, June 2022), p. 1, https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy.

82 Alice Saltini, "British thinking of AI integration into and interaction with nuclear command and control, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

83 *Artificial Intelligence in Support of Defence: Report of the AI Task Force* (Paris: French Ministry of the Armed Forces, September 2019), p. 5, https://www.defense.gouv.fr/sites/default/files/aid/Report%20of%20the%20AI%20Task%20Force%20September%202019.pdf.

84 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

85 Fei Su & Jingdong Juan, "Chinese thinking on AI integration and interaction with nuclear C2, force structure, and decisionmaking", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

86 Fei Su & Jingdong Juan, "Chinese thinking on AI integration and interaction with nuclear C2, force structure, and decisionmaking", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

87 Oleg Shakirov, "Russian thinking of AI integration into and interaction with nuclear C2, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

88 *2022 Nuclear Posture Review* (Washington D.C.: U.S. Department of Defense, 2022), p. 22, https://fas.org/wp-content/uploads/2023/07/2022-Nuclear-Posture-Review.pdf.

89 Marina Favaro, Neil Renic & Ulrich Kühn, " Negative Multiplicity: Forecasting the Future Impact of Emerging Technologies on International Stability and Human Security", *Institute for Peace Research and Security Policy* (September 2022), p. 66, https://ifsh.de/file/publication/Research_Report/010/Research_Report_010.pdf.

90 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

91 *Artificial Intelligence in Support of Defence: Report of the AI Task Force* (Paris: French Ministry of the Armed Forces, September 2019), p. 9, https://www.defense.gouv.fr/sites/default/files/aid/Report%20of%20the%20AI%20Task%20Force%20September%202019.pdf.

92 Fei Su & Jingdong Juan, "Chinese thinking on AI integration and interaction with nuclear C2,

force structure, and decisionmaking", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

93 Oleg Shakirov, "Russian thinking of AI integration into and interaction with nuclear C2, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

94 Tim McDonnell, Mary Chesnut, Tim Ditter, Anya Fink & Larry Lewis, "AI Implementation Matrix" in: *Artificial Intelligence in Nuclear Operations: Challenges, opportunities, and impacts, Center for Naval Analyses* (April 2023), https://www.cna.org/reports/2023/04/Artificial-Intelligence-in-Nuclear-Operations.pdf.

95 Fei Su & Jingdong Juan, "Chinese thinking on AI integration and interaction with nuclear C2, force structure, and decisionmaking", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

96 Ibid.

97 Ibid.

98 Elsa B. Kania, "CHINA'S RISE IN ARTIFICIAL INTELLIGENCE AND FUTURE MILITARY CAPABILITIES", *Center for a New American Security* (2017), https://www.jstor.org/stable/pdf/resrep16985.6.pdf

99 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

100 Ibid.

101 Ibid.

102 Tim McDonnell, Mary Chesnut, Tim Ditter, Anya Fink & Larry Lewis, "AI Implementation Matrix" in: *Artificial Intelligence in Nuclear Operations: Challenges, opportunities, and impacts, Center for Naval Analyses* (April 2023), p.9, https://www.cna.org/reports/2023/04/Artificial-Intelligence-in-Nuclear-Operations.pdf.

103 RESPONSIBLE ARTIFICIAL INTELLIGENCE STRATEGY AND IMPLEMENTATION PATHWAY (Washington D.C.: U.S. Department of Defense, June 2022), foreword, https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf.

104 2022 National Defense Strategy of The United States of America (Washington D.C.: October, 2022), p.6, https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF.

105 In his remarks for the Arms Control Association (ACA) Annual Forum, National Security Advisor Jake Sullivan claimed: "The P5 provides an opportunity manage nuclear risk and arms race pressures through a mix of dialogue, transparency, and agreements. For example, formalizing a missile launch notification regime across the P5 is a straightforward measure that is simply common sense. It's a small step that would help reduce the risk of misperception and miscalculation in times of crisis. And one that could potentially build momentum toward further measures to manage nuclear risks and arms racing— From maintaining a "human-in-the-loop" for command, control, and employment of nuclear weapons". See "Remarks by National Security Advisor Jake Sullivan for the Arms Control Association (ACA) Annual Forum | The White House", The White House https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/06/02/remarks-by-national-security-advisor-jake-sullivan-for-the-arms-control-association-aca-annual-for.

106 2022 National Defense Strategy of The United States of America (Washington D.C.: October, 2022), p.6, https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF.

107 Principles and responsible practices for Nuclear-weapon states: Working paper submitted by France, the United Kingdom of Great Britain and Northern Ireland and the United States of America (New York: 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, July 2022), https://www.un.org/sites/un2.un.org/files/npt_conf.2020_e_wp.70.pdf.

108 Historically, the Soviets considered developing a fully automated system that would transfer decision-making powers to machines during situations like a nuclear attack on Moscow or communication breakdown. This concept was ultimately discarded. However, they did develop a semi-automated system, known as the "Dead Hand", which could pass launch authority to ground commanders if an attack was detected. Detection was based on environmental cues like light radioactivity, seismic activity, and atmospheric pressure changes. See: *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy.* On the value of human judgement see Samuel Bendett, "Russian military debates AI development and use" in: *Strategic Insight 011/23*, The Azure Forum, (May 2023), https://www.azureforum.org/russian-military-debates-ai-development-and-use/.

109 Sari Arho Havrén, "China's No First Use of Nuclear Weapons Policy: Change or False Alarm?", *Royal United Services Institute* (October 2023), https://www.rusi.org/explore-our-research/publications/commentary/chinas-no-first-use-nuclear-weapons-policy-change-or-false-alarm#:~:text=In%20August%202023%2C%20at%20the,use%20nuclear%20weapons%20against%20non%2D ; Fiona S. Cunningham, "The Unknowns About China's Nuclear Modernization Program" *Arms Control Today*, (June 2023) https://www.armscontrol.org/act/2023-06/features/unknowns-about-chinas-nuclear-modernization-program ; Tong Zhao, "China and the international debate on no first use of nuclear weapons" in: *Asian Security* (December 2021) https://www.tandfonline.com/doi/abs/10.1080/14799855.2021.2015654?src=.

110 Fiona S. Cunningham, "The Unknowns About China's Nuclear Modernization Program" *Arms Control Today*, (June 2023), https://www.armscontrol.org/act/2023-06/features/unknowns-about-chinas-nuclear-modernization-program.

111 Margarita Konaev and Samuel Bendett, "Russian AI-Enabled Combat: Coming to a City near you?" *War on the Rocks* (July 2019), https://warontherocks.com/2019/07/russian-ai-enabled-combat-coming-to-a-city-near-you/ ; Alexa Wehsener, Andrew W. Reddie, Leah Walker, Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures, *Institute for Security and Technology* (February 2023), p. 10, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

112 Lydia Wachs, "The Role of Nuclear Weapons in Russia's Strategic Deterrence: Implications for European security and nuclear arms control", *SWP Comment NO.68*, (November 2022), p.1, https://www.swp-berlin.org/10.18449/2022C68/.

113 Alexa Wehsener, Andrew W. Reddie, Leah Walker, Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures, *Institute for Security and Technology* (February 2023), p. 15-16, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

114 2022 Nuclear Posture Review (Washington D.C.: U.S. Department of Defense, 2022), p. 8, https://fas.org/wp-content/uploads/2023/07/2022-Nuclear-Posture-Review.pdf.

115 *2022 Nuclear Posture Review* (Washington D.C.: U.S. Department of Defense, 2022), p. 9, https://fas.org/wp-content/uploads/2023/07/2022-Nuclear-Posture-Review.pdf.

116National Strategic Review 2022 (Paris: Secrétariat général de la défense et de la sécurité nationale, 2022), https://www.sgdsn.gouv.fr/files/files/rns-uk-20221202.pdf.

117 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

118 Claire Mills, *Nuclear weapons at a glance: United Kingdom*, House of Commons Library, (May 2023), https://researchbriefings.files.parliament.uk/documents/CBP-9077/CBP-9077.pdf.

119 Alice Saltini, "British thinking of AI integration into and interaction with nuclear command and control, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

120 D*OD DIRECTIVE 3000.09 AUTONOMY IN WEAPON SYSTEMS*, Office of the Under Secretary of Defense for Policy, (January 2023), https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DOD-DIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF.

121 Defence Artificial Intelligence Strategy (London: UK Ministery of Defence, June 2022), p. 59, https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy.

122 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", European Leadership Network (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

123 *Principles and responsible practices for Nuclear-weapon states: Working paper submitted by France, the United Kingdom of Great Britain and Northern Ireland and the United States of America* (New York: 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, July 2022), https://www.un.org/sites/un2.un.org/files/npt_conf.2020_e_wp.70.pdf.

124 Michael Kofman, Richard Connolly, Jeffrey Edmonds, Andrea Kendall-Taylor, & Samuel Bendett, "Assessing Russian State Capacity to Develop and Deploy Advanced Military Technology", *Center for New American Security* (October 2022), p. 1, https://www.cnas.org/publications/reports/assessing-russian-state-capacity-to-develop-and-deploy-advanced-military-technology#:~:text=The%20report%20identifies%20two%20drivers,force%20choices%20on%20the.

125 A*I Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Washington D.C.: U.S. Defense Innovation Board, 2019) https://innovation.defense.gov/ai/.

126 Ambitious, Safe, Responsible: Our approach to the delivery of Alenabled capability in Defence (London: Ministry of Defence, June 2022), p. 12, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082991/20220614-Ambitious_Safe_and_Responsible.pdf.

127 Héloïse Fayet, "French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

128 Defence Artificial Intelligence Strategy (London: UK Ministery of Defence, June 2022), p. 59, https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy

129 United Nations General Assembly, "Without Adequate Guardrails, Artificial Intelligence Threatens Global Security in Evolution from Algorithms to Armaments, Speaker Tells First Committee", First Committee, Seventy-Eight Session (New York, 24 October 2023), https://press.un.org/en/2023/gadis3725.doc.htm.

130 "Position Paper of the People's Republic of China on Regulating Military Applications of Artificial Intelligence (AI)", PERMANENT MISSION OF THE PEOPLE'S REPUBLIC OF CHINA TO THE UNITED NATIONS OFFICE AT GENEVA AND OTHER INTERNATIONAL ORGANIZATIONS IN SWITZERLAND (December 2021), http://geneva.china-mission.gov.cn/eng/dbdt/202112/t20211213_10467517.htm.

131 "Secretary-General's remarks at press conference launching High-Level Advisory Body on Artificial Intelligence | 26 October 2023, New York", The United Nations, https://www.un.org/sg/en/content/sg/press-encounter/2023-10-26/secretary-general%E2%80%99s-remarks-press-conference-launching-high-level-advisory-body-artificial-intelligence%C2%A0.

132 Oleg Shakirov, "Russian thinking of AI integration into and interaction with nuclear C2, force structure and decision-making", *European Leadership Network* (November 2023), https://www.europeanleadershipnetwork.org/ai-enabled-decision-making/.

133 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures", *Institute for Security and Technology* (February 2023), p.30, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

134 Ibid, pg 27

135 "FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence", The White House, (October 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

136 The alignment problem refers to the challenge of ensuring that the goals and behaviours of AI systems align with human values and intentions.

137 Jared Kaplan, Sam McCandish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei, "Scaling Laws for Neural Language Models", *Open AI* (January 2023), https://arxiv.org/pdf/2001.08361.pdf

138 Alice Saltini, "To Avoid Nuclear Instability, a Moratorium on Integrating AI into Nuclear Decision-making is Urgently Needed. The NPT PrepCom Can Serve as a Springboard", *European Leadership Network* (28 July 2023), https://www.europeanleadershipnetwork.org/commentary/to-avoid-nuclear-instability-a-moratorium-on-integrating-ai-into-nuclear-decision-making-is-urgently-needed-the-npt-prepcom-can-serve-as-a-spring-board/.

139 "Joint Statement of the Leaders of the Five Nuclear-Weapon States on Preventing Nuclear War and Avoiding Arms Races" (Washington DC: The White House, 3 January 2022), https://www.whitehouse.gov/briefing-room/statements-releases/2022/01/03/p5-statement-on-preventing-nuclear-war-and-avoiding-arms-races/.

140 Alexa Wehsener, Andrew W. Reddie, Leah Walker, & Philip J. Reiner, "AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures", *Institute for Security and Technology* (February 2023), p.22-23, https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf.

141 Haroon Sheikh, Corien Prins & Erik Schrijvers, "Artificial Intelligence: Definition and Background" in: *Mission AI* (Cham: Springer, 2023), pp.15 – 41, https://doi.org/10.1007/978-3-031-21448-6_2.

142 Laith Alzubaidi, Jinglan Zhang, Anjad J. Humaidi, et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions", *Journal of Big Data* 8, 53 (2021), https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8.

143 Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models", *arXiv* (12 July 2022), https://arxiv.org/abs/2108.07258v3.

144 Russ Altman, Erik Brynjolfsson, Michele Elam, Surya Ganguli, Daniel E. Ho, James Landay, Curt Langlotz, Fei-Fei Li, Percy Liang, Christopher Manning, Peter Norvig, Rob Reich, Vanessa Parli: "Generative AI: Perspectives from Stanford HAI", *Stanford University Human-Centered Artificial Intelligence* (March 2023), https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives.pdf.

145 "Better language models and their implications", *OpenAI* (14 Febrary 2019), https://openai.com/research/better-language-models ; Samuel R. Bowman, "Eight Things to Know about Large Language Models", *arXiv* (April 2023), https://arxiv.org/pdf/2304.00612.pdf.

146 "What is machine learning?", IBM, https://www.ibm.com/topics/machine-learning, ; Batta Mahesh, "Machine Learning Algorithms -A Review", *International Journal of Science and Research* (January 2019), https://www.researchgate.net/profile/Batta-Mahesh/publication/344717762_Machine_Learning_Algorithms_-A_Review/links/5f8b2365299bf1b53e2d243a/Machine-Learning-Algorithms-A-Review.pdf?eid=5082902844932096.

147 Yuanyuan Tian, Mi Shu & Qingren Jia, "Artificial Neural Network" in *Encyclopedia of Mathematical Geosciences*, ed. by Daya Sagar, Qiuming Cheng, Jennifer McKinley, Frits Agterberg (Cham: Springer, 2021), https://doi.org/10.1007/978-3-030-26050-7_44-1. "Artificial Neural Networks: What they are & why they matter", *SAS*, https://www.sas.com/en_us/insights/analytics/neural-networks.html.

148 Haroon Sheikh, Corien Prins & Erik Schrijvers, "Artificial Intelligence: Definition and Background" in: *Mission AI* (Cham: Springer, 2023), pp.15 – 41, https://doi.org/10.1007/978-3-031-21448-6_2.

149 Batta Mahesh, "Machine Learning Algorithms -A Review", *International Journal of Science and Research* (January 2019), https://www.researchgate.net/profile/Batta-Mahesh/publication/344717762_Machine_Learning_Algorithms_-A_Review/links/5f8b2365299bf1b53e2d243a/Machine-Learning-Algorithms-A-Review.pdf?eid=5082902844932096.

150 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", *arXiv* (August 2023), https://arxiv.org/pdf/1706.03762.pdf ; Rick Merritt, "What Is a Transformer Model?", *NVIDIA* (25 March 2022), https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model)/.

151 Haroon Sheikh, Corien Prins & Erik Schrijvers, "Artificial Intelligence: Definition and Background" in: Mission AI (Cham: Springer, 2023), pp.15 – 41, https://doi.org/10.1007/978-3-031-21448-6_2.

The European Leadership Network (ELN) is an independent, non-partisan, pan-European NGO with a network of over 300 past, present and future European leaders working to provide practical real-world solutions to political and security challenges.

About the author

Alice Saltini
Research Coordinator, European Leadership Network

Contact

Published by the European Leadership Network, November 2023

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

@theELN | europeanleadershipnetwork.org

The opinions articulated in this report represent the views of the author, and do not necessarily reflect the position of the European Leadership Network or any of its members. The ELN's aim is to encourage debates that will help develop Europe's capacity to address pressing foreign, defence, and security challenges.

EUROPEAN LEADERSHIP NETWORK

Scan to find out more about the project and to read the Chinese, Russian, French and UK bibliographies:

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