Framework for Responsible Use of AI in the Nuclear Domain

This policy brief outlines the need for an international framework addressing the convergence of artificial intelligence (AI) and nuclear command, control, and communications (NC3) systems.

5th February 2025

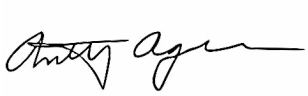
Policy Team

policy@futureoflife.org

Contents

Preface

As artificial intelligence rapidly advances in capability as well as autonomy, humanity faces a fundamental question: which of our many decisions and institutions must we keep firmly under meaningful human control? None could be more of clear existential importance than the control of nuclear weapons. There is agreement among most world leaders and experts that the AI finger should be nowhere near the nuclear button; we can hope that our leaders can formalize this consensus. But thorny questions remain about AI's involvement in the decision process and in how crucial information is aggregated and provided to leaders for decision. Multilateral discussions such as those described here are crucial in building common expert and political understanding of how to ensure that nuclear weapons remain controlled, secure, and—we may all hope—unused.



Anthony Aguirre

Executive Director,

Future of Life Institute

Introduction

From January 2022 to December 2024, the Geneva Centre for Security Policy (GCSP) and Strategic Foresight Group (SFG) have facilitated an experts' dialogue process on nuclear risk reduction between five permanent members of the UN Security Council – China, France, Russia, the United Kingdom, and the United States – known as the Normandy P5 Initiative. It was supported by the Future of Life Institute (FLI) and the Silicon Valley Community Foundation, with occasional extra support from the Swiss Federal Department of Foreign Affairs and Normandy for Peace Initiative of Region Normandy. The initiative is inspired by the Normandy Manifesto for World Peace, issued by a group of Nobel Peace Laureates and social thinkers, in June 2019, warning the world about existential risks to human civilization posed by weapons of mass destruction and the growing influence of new technologies on matters determining the fate of humanity. The Normandy P5 Initiative began with broad concern for the implications of catastrophic risks for global security and has gradually focussed on the application of AI and other technologies in nuclear command, control and communications including decision support infrastructure. This process involved four in-person roundtables, several online conferences, bilateral consultations with the P5 Disarmament Ambassadors and other stakeholders.

The concluding roundtable was hosted in Geneva from 4 to 6 December 2024 with the participation of experts from the P5 countries who included political leaders, military officials, specialists in AI and academic scholars.

Rationale

Since early 2022, when the initiative was launched, the geopolitical environment has deteriorated significantly, interrupting channels of communication between the five countries.

In addition to or because of geopolitical tensions, the last three years have seen an increasing tendency to modernize nuclear weapons. Experts worry about the temptation of the nuclear power states to target each other's nuclear forces and command and control for pre-emptive attacks to limit the damage they could do in response. It is this attempt to negate or severely degrade the other side's nuclear deterrent that creates the great time pressure to launch, resulting in crisis escalation and instability. AI could relieve or exacerbate these pressures.

During this period of deteriorating environment, some discussions have been taking place between like-minded countries on nuclear risks and separately on the military uses of AI, but not on the linkages between the two domains. An official P5 nuclear risk reduction dialogue between experts and young scholars which was initiated in 2009 is continuing under the joint sponsorship of the five governments. It has its origins in the nuclear non-proliferation agenda. This process does not include discussion on issues related to the interface between AI and nuclear weapons.

In 2024, the responsible use of AI in military in general and in the nuclear domain in particular, was discussed in some intergovernmental fora. In September 2024, REAIM conference in Seoul affirmed the principle of human control in the interface between AI and NC3. This statement was signed by 60 countries including 3 permanent members of the Security Council (US, UK, France). In November 2024, the Presidents of the People's Republic of China and the United States affirmed the need to address the risk of AI systems, improve AI safety and international cooperation and to maintain human control over the decision to use nuclear weapons. It is not clear if the appreciation of the risk emanating from AI-nuclear convergence reflected in some of the intergovernmental statements in 2024 will strengthen or weaken in 2025, as new governments take office in some parts of the world.

In view of these uncertainties, and considering that AI systems are not predictable, reliable, or verifiable, it is necessary to advocate an international framework for the responsible use of AI in the nuclear domain, which factors in unpredictable future risks. It is necessary to recognise that an agreement on such a framework would not be easy, as adversaries may find it difficult to agree on a common ground for the following reasons:

- National security concerns in conflict with the principle of transparency
- Lack of internal clarity in each of the P5 countries on how AI might be integrated into nuclear command
- Lack of clarity about the use of AI over and above automation
- Lack of mutual trust in a complex geopolitical context
- Competitive dynamics and the fear of losing strategic edge
- Divergence on legal frameworks, ethical standards, and cultural perspectives
- Classified nature of information related to AI algorithms and NC3 systems
- Cybersecurity risks
- Dual use nature of many AI applications
- Denial of risks.

It is possible that a state may want to deny the risks emanating from AI-NC3 convergence to assuage public fear, at a time when the state wants to invest heavily in these technologies and needs to mobilise the support of its population.

Risks

Even though a framework agreed by the P5 governments for responsible AI-NC3 is difficult for the reasons mentioned above, it is essential. In the absence of such a framework, the AI-NC3 interface can pose unintended consequences, and catastrophic risks to humanity, and not merely to the countries engaged in an adversarial relationship. The fog of war is likely to get much worse as AI contributes heavily to information warfare. Some of the risks include:

- False alarms or signals misinterpretation or false positives (unintended) – identifying an incorrect or non-existing nuclear threat leading to a cascading series of events
- Excessive autonomy and lack of human oversight (unintended) – can lead to AI algorithms generating a process that results in the launching of nuclear weapons
- Inability of AI to distinguish between military and civilian targets due to manipulation of training datasets
- Absence or failure of crisis communication channels
- An uncomfortably high margin of error. The experience with the use of AI in non-military applications such as crime detection and facial recognition software has shown a margin of error that is unacceptable
- Rushing into development of AI-NC3 capabilities that have not been properly tested and/or verified.

Risks related to concurrent developments in AI and Cyber include:

- Cyberattacks (intended) – hacking and unauthorised access leading to wrong commands to deploy nuclear weapons
- Spear phishing (manipulation of principals controlling the nuclear chain) to get access to sensitive information.

AI in nuclear domain can be a double-edged sword. It can help improve accuracy in threat detection or target detection. It can however, fail due to (1) malfunctioning or (2) manipulation, resulting in catastrophic consequences.

The main function of AI in NC3 is threat detection. This is achieved in the following ways:

- AI technologies can be used to process and analyse vast amounts of data generated by the various sensors, satellites, and communication systems that are part of the NC3 network. This can include:
  - Signal Processing: Enhancing the detection and interpretation of signals amidst noise, which is crucial for early warning systems
  - Data Fusion: Combining data from multiple sources to create a more accurate and comprehensive understanding of the nuclear landscape.
- AI can be deployed to:
  - Monitor networks for unusual activities indicating cyber threats
  - Automate responses to identified cyber threats
  - Predict potential vulnerabilities and suggest countermeasures.
- AI can underpin decision support systems that assist human operators in making informed decisions. These systems can:
  - Provide real-time analytics and recommendations
  - Highlight potential consequences of different courses of action
  - Enhance the speed and accuracy of decision-making under stress.

While on the one hand, the threat detection functions of AI outlined above can help overcome human limitations, unintentionally wrong threat detection can trigger an attack against a non-existing perceived threat setting events into motion that result in a nuclear war. Also, reducing decision making time due to high speed of AI aided analysis has double-edged consequences. If one party has higher speed, it is an advantage that gives them more time to make good decisions. If both sides optimize for speed, there is no advantage, but more decisions may have to be effectively delegated to time-saving technology such as AI.

However, AI analysis could extend human decision-making time: if an AI analysis requires less time for a task than a human analyst performing the same task, it could mean that the person assessing the analysis has more time to consider the outcome.

AI has an impact on the psychology of decision-making. The speeding up of the decision loop (the time leaders would have to decide whether to use a nuclear weapon) poses a critical problem. Even if it is universally agreed not to permit AI in the nuclear command functions, retaining the human in the loop, the functionaries that brief a leader would use AI as a decision support tool to gather their input at an extremely fast speed. This can expose them to the risk of unintentional wrong threat detection. In addition, there is evidence that users of AI enabled systems can become over-reliant and over-trusting of AI and not critique or over-ride the decisions an AI-enabled system makes.

Related to this is the issue of AI and the information environment, particularly the lack of correct information about what the adversaries are doing in this area. This is bound to cause uncertainty leading to wrong hypothesis. To overcome this, it would be useful to understand how the AI systems of the adversary are built. However,

that would require making information publicly available. At a time of heightened tensions and competition, the incentives for doing this are low.

A nuclear war can take place due to intent, incident or accident. There may not be an intent to wage a nuclear war, and the warring parties may want to confine to a conventional war. But a conventional war can escalate into a nuclear war irrespective of the extent to which AI is integrated in NC3.

AI can be useful in simulating models for nuclear conflict. However, accurate AI models essentially depend on vast amount of data. Since there has not been any real nuclear conflict since 1945, simulation based on purely mathematical hypotheses or synthetic data can lead to wrong calculations which could have disastrous consequences. For instance, if AI systems are confronted with real-world scenarios that are different from the data they have been trained on, they may malfunction in the sense that they have not stopped working but rather produce unpredictable and unexplainable results.

The increasing use of AI for intelligence, surveillance, and reconnaissance (ISR) can have an impact on postures and strategic stability. This applies essentially to China, which relies on mobile launch systems (that move around on land) - their strategic deterrence relies on the US not 'knowing' where all its nukes are, giving second strike capability. This is less true for the US, France, Russia and UK because they currently rely much more on submarine systems. It is also unclear if the breakthrough in ISR would come from AI or from corollary technologies, such as quantum computing and imaging. If one country perceived that AI would give its adversary the first mover advantage because it had enabled them to find all its nuclear weapons, then that would undermine its deterrent. It could also lead to nuclear weapons being used earlier in a crisis than they would otherwise. Also, such a race creates greater incentives for expanding the arsenals.

The threats to associated infrastructure, such as fibre optics and satellites, also need to be considered. Interference in these communications channels can impact early warning systems.

Framework

To mitigate the risks such as the ones mentioned above, it is necessary to develop a framework for the responsible use of AI and other emerging technologies in the nuclear domain. The primary political challenge remains a return to global governance through multilateralism and cooperation, which have been a pillar of world peace since 1945. The initiative for such a framework should be driven by all P5 counties in an inclusive manner, without excluding any of the P5 countries from the process. Below are elements of a potential framework for cooperation between the P5 countries.

Principles of Responsible AI use in NC3

Such a framework can be based on the following principles:

- Transparency and explainability – clear understanding about how algorithms operate, and decisions made by AI should be explainable to human operators. Although it may be utopian to demand any form of transparency, especially in AI algorithms. These tools are highly sensitive and surrounded by rules of confidentiality and secrecy that are difficult to be shared between the P5 nations, which are commercial competitors, technological rivals and sometimes adversaries in latent conflicts or international crises. It will be particularly difficult to institute verification measures, as AI algorithms are invisible, unlike verification measures for nuclear missiles and facilities. However, to begin with, the principle of explainability should be promoted while the principle of transparency can be pursued in a distant future.
- Human control – use AI only as a decision support tool and not as a replacement for human judgement; avoidance of autonomous weapons systems in the context of NC3 and its delivery systems ensuring human operators retain control on initiation and execution of all nuclear decisions - including assessment of inputs from AI decision support systems. If probabilistic AI is used for threat detection or other decision support systems, there should be another source of input for comparative assessment. To ensure the implementation of this principle, it would be necessary to airgap the launch command from the decision support and early warning systems. There would be a risk of misjudgement if there is very little time between the input of threat data and the need to take a launch decision. It is not known whether in future the algorithms would even allow effective human control.
- Adherence to international humanitarian law – no targeting of civilian population, cultural centres, and non-military infrastructure. While IHL is meant to be respected even in a conventional war, irrespective of technology used, it is important to underline its significance in cases where AI is used.

Prohibition on Offensive AI Capacities

Parties undertake not to develop, deploy, or use AI applications in NC3 systems with the intent to cause harm, disruption, or unauthorized access to nuclear capabilities of other parties.

- Prevent malicious manipulation of data – not to employ AI algorithms for manipulating data in NC3 systems about status or readiness of nuclear forces
- Ensure no deliberate creation and distribution of poisoned datasets
- No cyberattacks on NC3 systems and critical crisis communication channels
- Avoid the development or use of AI systems that might autonomously initiate pre-emptive strikes based on predictive algorithms, recognising the risk of miscalculations and unintended consequences.

National Level Voluntary Measures

- Regular national level inspections and audits of AI systems used in NC3 to ensure compliance with agreed principles and safety standards
- Research into the development of fail-safes (such as redundant algorithms meant to do the same task but trained with different datasets)
- Maintain and improve air gapping of crisis communication channels
- Robust cybersecurity and other measures to protect AI systems from unauthorised access, hacking, and manipulation; also prevent spoofing and jamming of signals by adversaries with anti-jamming and signal authentication technologies
- Practise crisis management, including war games and crisis simulations involving AI integration into NC3
- Establishment of Ethics Committees and Safety Institutes within national jurisdiction of each of the P5 countries.

International Collaboration

- Expansion of the nuclear glossary already prepared by the P5 governmental experts process to include the AI dimension
- All P5 countries should publish their nuclear doctrines, including the dimension of convergence between AI and NC3
- Joint guidelines for the development, testing, deployment of AI in the nuclear domain with safety and ethical consideration, giving priority to the protection of human life, human rights, and international humanitarian law
- Development of new, transparent, international standards for robust test, evaluation, validation and verification (TEVV) protocols for AI in NC3, including international agreement on performance standards
- Joint research in collaborative solutions to potential problems
- Crisis communication mechanism to establish quick communication in case of malfunctioning of AI systems in NC3 or in case of any doubt. (This can be developed with provably secure software.) It is necessary to examine whether this would require new mechanisms or improvisation of some of the existing systems
- Joint training and capacity building programmes to enhance understanding of AI ethics, safety issues, best practices and simulation exercises, including the training of developers and operators of AI enabled weapons systems in international humanitarian law
- Periodic meetings at the technical level on the adherence to principles, political declarations, and agreements if any (at technical level)
- A meeting once in two years of at the level of senior military leaders and other functionaries to discuss AI-nuclear concerns
- P5 joint guidelines to prevent the development and use of AI systems that might automatically initiate pre-emptive strikes based on predictive algorithms
- Creation of an intergovernmental agency for the governance of AI in NC3 to recommend standards, suggest common evaluation and testing metrics, conduct training programmes, and convene summit level meetings. Such an agency may be created by P5 countries and eventually be expanded as an organisation for global governance. If there is political opposition to such a proposal, it may be developed in phases with the initial mandate limited to conduct joint training and capacity building exercises.

High Level Political Declaration

A high-level political declaration jointly by the P5 countries committing to retain human responsibility over nuclear decision-making, including the decision support and communication systems, is needed to provide strategic direction to the discourse on the interface between AI and the nuclear domain. It should be achievable as the five have already made such statements individually or in a small group. The P5 leaders issued a joint statement committing themselves to avoid nuclear war on 3 January 2022. A political declaration on human primacy over AI in NC3 can be conceived as a follow-up to the joint statement on the avoidance of nuclear war.

Such a statement can endorse the principles, national level voluntary measures and an architecture for international cooperation outlined above. The preparatory discussions can include an exchange on what assurances the P5 can give each other over their use of AI in nuclear systems. They can also discuss their threat perceptions and their fears over how the technology might be used. Such exchanges can lead to a common understanding on not using the most dangerous systems.

The process for issuing a high-level political declaration can be bottom-up or top-down. In the bottom-up approach, the official P5 dialogue on nuclear risk reduction can include the interface between AI and NC3 on their agenda on a sustained basis with a discussion on the issues identified above, among others.

In the top-down approach, the Heads of Government of the P5 countries, when politically and strategically feasible, may issue a short statement confirming the primacy of the human in the nuclear domain and authorise their governments to prepare the framework for the AI and nuclear domain interface, through technical, legal, and political exchanges at the senior level of governments.

The political declaration can be further enhanced by its endorsement by the UN Security Council.

Conclusion

Once there is an agreement between P5 countries on such a framework, it would be important to expand its scope to cover all nuclear states and eventually into an instrument of global multilateral governance, including the prevention of misuse of AI by terrorist groups.

This initiative is inspired by the Normandy Manifesto for World Peace of 2019 which calls for phased elimination of all weapons of mass destruction. An agreement on a framework between the P5 countries on AI-nuclear interface is seen as the first step to serve the long-term vision of a world free of nuclear weapons.

About the Co-Authors

SFG and FLI are grateful to the Geneva Centre for Security Policy (GCSP) for hosting several roundtables in Geneva, which contributed to this paper.

Strategic Foresight Group

Strategic Foresight Group is an international think-tank based in Mumbai, India. It has worked with governments and national institutions in 65 countries, since its inception in 2002. It helps policy makers to anticipate global challenges such as catastrophic wars, transformative technologies, cross-boundary water conflicts and terrorism. It develops new policy concepts and convenes leaders from rival countries to craft collaborative solutions. The research output and policy prescriptions of SFG have been discussed in the United Nations Security Council, UN Alliance of Civilizations, World Economic Forum at Davos, Interaction Council of former Heads of States, European Parliament, Indian Parliament, UK House of Lords, House of Commons and other institutions.

About Future of Life Institute

The Future of Life Institute (FLI) is an independent nonprofit organization with the goal of reducing large-scale risks and steering transformative technologies to benefit humanity, with a particular focus on artificial intelligence (AI). Since its founding, FLI has taken a leading role in advancing key disciplines such as AI governance, AI safety, and trustworthy and responsible AI, and is widely considered to be among the first civil society actors focused on these issues. FLI was responsible for convening the first major conference on AI safety in Puerto Rico in 2015, and for publishing the Asilomar AI principles, one of the earliest and most influential frameworks for the governance of artificial intelligence, in 2017. FLI is the UN Secretary General's designated civil society organization for recommendations on the governance of AI and has played a central role in deliberations regarding the EU AI Act's treatment of risks from AI. FLI has also worked actively within the United States on legislation and executive directives concerning AI. Members of our team have contributed extensive feedback to the development of the NIST AI Risk Management Framework, testified at Senate AI Insight Forums, briefed the House AI Task-force, participated in the UK AI Safety Summit, and connected leading experts in the policy and technical domains to policymakers across the US government.