July 2023

# Risks of Artificial Intelligence in Nuclear Command, Control and Communications (NC3)

*Primer & Policy Options for Risk Mitigation*

**Contact**
policy@futureoflife.org

The aim of this brief is to serve as a primer for those interested in what has been written about the risks of AI integration into NC3 and the potential mitigation measures proposed within that literature. This is intended as a review, and is therefore not reflective of the Future of Life Institute (FLI)'s policy positions.

High-stakes applications of AI, such as those in military domains, warrant careful consideration, testing and analysis to ensure that the risks of adopting AI do not outweigh the potential benefits. Given the potentially catastrophic costs of failure, military applications require exceptional prudence. There has been international dialogue and research on the application of artificial intelligence in weapons systems; however, the international forum in which this issues has been discussed, the UN CCW, has been limited by its mandate only to consider the risks of conventional weapons systems. The International Committee for the Red Cross has formally proposed a policy architecture of prohibitions and regulations to mitigate the risks of autonomy in weapons systems. Even if such a proposal were to be adopted within the UN CCW, the policy would be limited to conventional weapons.

FLI has argued that integration of AI into weapons with the capability of mass destruction, such as nuclear, chemical, and biological weapons poses substantial risks to society. These risks are further augmented when human control is eroded or removed. There is increasing interest by states to integrate AI throughout conventional and nuclear command, control and communications (NC3). The effect of AI integration on strategic stability and nuclear risk has received relatively limited exploration, research, or international dialogue. The paucity of work in this area is set against a backdrop of prevailing and increasing arms race dynamics between the major technology and nuclear military powers. In fact, the geopolitical environment has led many to conclude that integration of AI into NC3 is inevitable.

**Thorough exploration of strategic stability and nuclear risk in the context of AI is urgently needed to identify where to draw lines on acceptable and unacceptable applications, develop robust mitigation measures, and identify stabilizing policies to prevent intentional or accidental nuclear conflict.**

For those interested in a detailed dive into this topic we recommend reading:

1. Boulanin V, Saalman L, Topychkanow P, Su F, Carlsson MP. *Artificial Intelligence, Strategic Stability and Nuclear Risk.* Stockholm International Peace Research Institute. 2020.
2. Hruby J, Miller MN. *Assessing and Managing the Benefits and Risks of Artificial*

*Intelligence in Nuclear-Weapon Systems*. Nuclear Threat Initiative. 2021.

3. Wehsener A, Walker L, Beck R, Philips L, Leader A, *Forecasting the AI and Nuclear Landscape*, Institute for Security and Technology. September 2022.

# Part 1: Risks of Artificial Intelligence Integration to Nuclear Command, Control, Communications

Artificial intelligence can source massive amounts of data across a complex network, process that data at machine speed, and then output a simple, understandable recommendation to human operators — or in some circumstances execute tasks without any human involvement. This power introduces a host of risks, the consequences of which are highly dependent on the particular application:

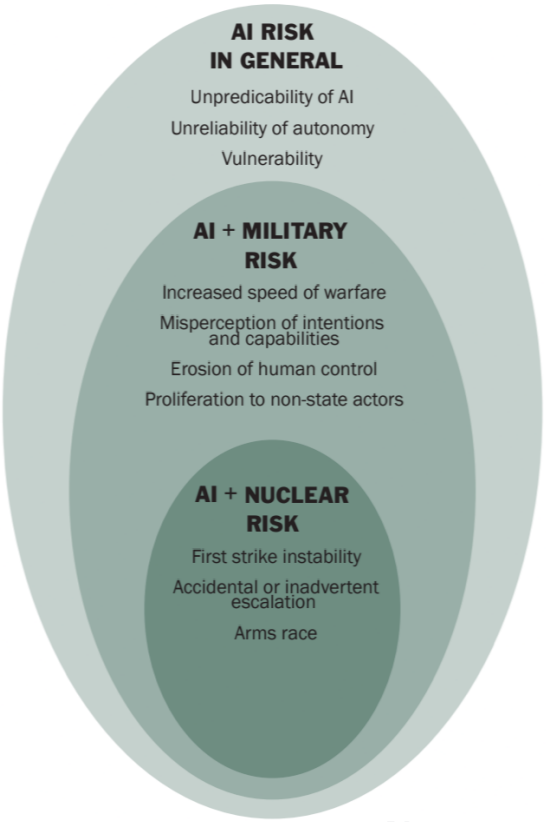



*Figure 1: Sourced from Boulanin V et al, depicts the layers of risk associated with AI.*

***Layer 1: "Risks inherent to the nature and limitations of AI technology.*** These include broad and general challenges such as the unpredictability of machine learning systems, the lack of reliability of autonomous systems, and their vulnerability to adversarial attack such as cyberattack, data poisoning and spoofing."

***Layer 2: "Risks posed by the use of AI for military applications:*** These range from the challenge of a state signaling its own capabilities and intentions and understanding those of its opponent. This is particularly the case when AI-powered military technologies are used to deal with the acceleration of the speed of warfare. A related risk in that regard is the potential erosion of human control over the use of force. A further key concern is the acquisition of military AI by non-state actors, which is facilitated by the dual-use nature of AI technology." We would add that when the critical functions of a weapon are performed by algorithms, this enables weapons systems to scale far beyond the number of human operators (e.g. # weapons / # human operators >> 1) - the "human-machine ratio" referenced in other SIPRI reporting on autonomy in weapons systems.

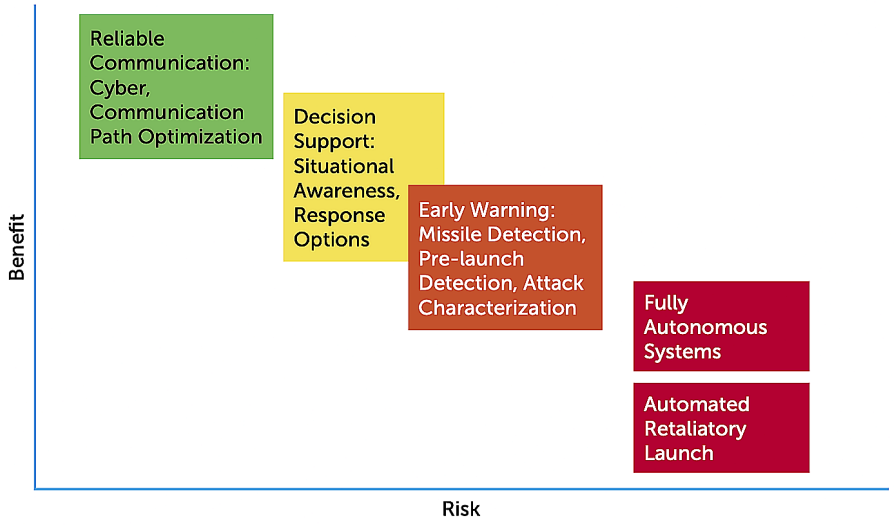
***Layer 3: "Risks posed by the use of AI in connection with nuclear weapon systems:*** These include AI undermining the confidence of nuclear-armed states in their second-strike capabilities. AI may also be employed to weaken the cybersecurity of nuclear force-related systems. AI also has the potential to provide new tools for influence operations on nuclear decision makers. It can increase the risk of accidental or inadvertent escalation, due to system failure or misuse of technology. Finally, there is the risk of deliberate escalation into nuclear conflict due to conventional force asymmetry fueled by AI advances."

This risk framework proposed by SIPRI is a concise yet reasonably comprehensive overview of the main technical risks associated with AI integration into nuclear weapons systems. It is worth noting that such technical risks can be further exacerbated or mitigated based on the environment and geopolitical context of deployment. Highly networked systems such as those of NC3 are more vulnerable than sets of isolated smaller systems, because in a highly networked system, failure at any point in the system can be amplified to imperil the network at large. Arms race dynamics and geopolitical tensions can favor escalation over de-escalation, speed over safety, and conflict over diplomacy, and increase the risk of misperceptions of intentions and capabilities. Such dynamics also incentivize rapid adoption and use of novel technical solutions, which may lead to the premature deployment of

technologies and develop a culture of automation bias. Further, it is also worth noting that given the nature of nuclear conflict, the data to train algorithms will largely need to be simulated. There are many documented examples of AI failing real-world deployment even when it is trained on real-world data, which raises significant concerns on the reliability of simulated data in such high-stakes decisions.

The Nuclear Threat Initiative aimed to stratify the potential risks of AI on nuclear stability by application area within NC3. Their analysis revealed that there may be net stabilizing applications of AI, especially in the area of improving communications. However, the vast majority of applications of AI in NC3 had an uncertain or net destabilizing effect on nuclear ability (Fig 2 ). This result raises serious concerns about broader integration of AI into NC3 increasing the risk of nuclear war.

**Figure 4: Relative Benefit, Risk, and Impact on Strategic Stability of AI Applications to Nuclear Weapons and Their Operational Systems**




Note: The impact on strategic stability is indicated by the color of the application box, with green being the most likely to have a stabilizing effect, yellow and orange indicate careful consideration for impact on stability, and red is most likely to have a destabilizing effect.

Further, while we cannot know for sure, the risk of AI integration into NC3 is likely material on short time horizons. The probability of AI integration occuring on short time scales, combined with the destabilizing, and potentially catastrophic effects, highlights the urgency of risk mitigation research.

# Part 2: Examples of Policy Options to Mitigate Risks

Integration of AI into NC3 is not inevitable; civil society and many states have the ability to influence this trajectory. As previously noted, there are applications of AI that may have net stabilizing effects, while others could be catastrophically destabilizing. Given the number of AI applications that are likely to raise serious concerns (Fig 2) or are clear to have a net destabilizing effect, it is important to consider a comprehensive range of policy options. Below we highlight some of the key options proposed by various analyses and agencies. This list is limited and non-exhaustive, and is meant to highlight select policy themes:

## Clear Policies on Where to Draw the Line

**Prohibitions and Regulations on AI Integration into NC3:** Prohibitions on the highest-risk applications of AI in NC3 could be considered. These could include prohibitions on fully-autonomous systems and on automated retaliatory launch. Strict regulations could be considered for the use of AI in early warning systems, missile detection, pre-launch detection, and attack characterization, as well as in decision support systems, situational awareness and the generation of response options.

**Meaningful human control –** A clear starting point for mitigating risk is to ensure that the decision to use nuclear weapons always be made by humans, and not algorithms. Critically, the role of humans in nuclear decision-making must be meaningful - that is, central to the selection of targets, to the time, location, and manner of use, and to the ultimate decision to deploy - rather than functioning only in an ancillary or oversight capacity relative to AI. This starting point is absolutely necessary, but in isolation, it is not sufficient to eliminate risk due to the complexity of the risk landscape detailed in Figure 1. Beyond the work of SIPRI and NTI, a key recommendation of the National Security Commission on Artificial Intelligence was for the United States government to *"Clearly and publicly affirm existing U.S. policy that only human beings can authorize employment of nuclear weapons, and seek similar commitments from Russia and China."* Further, this report goes on to note that *"Although joint political commitments that only humans will authorize employment of nuclear weapons would not be verifiable, they could still be stabilizing, responding to a classic prisoner's dilemma: as long as countries have confidence that others are not building risky command and control structures that have the potential to inadvertently trigger massive nuclear escalation, they would have less incentive to develop such systems themselves."* This finding is important, as it highlights that

while verifiable agreements are the ideal standard, agreements that are not technically verifiable can still do their part to mitigate risk. The United States recently reiterated this recommendation. Similar commitments have been detailed in the UK's Defense Artificial Intelligence Strategy, with a commitment to "*ensure that – regardless of any use of AI in our strategic systems – human political control of our nuclear weapons is maintained at all times.*" (SIPRI, NTI, NSCAI)

**No First Use policies –** Adoption of clear 'No First Use' policies by all nuclear-armed states is another possible measure. With 'No First Use' policies states commit never to use nuclear weapons first, either as a first strike or in retaliation to a conventional attack. Out of the nuclear armed states, only China and India have adopted this policy. (SIPRI)

**Commitment to lower the alert status of nuclear arsenals –** Removing strategic weapons from a launch-on-warning or launch-ready alert status might help to mitigate risk, too. One example of this is keeping nuclear warheads and delivery mechanisms separate, as is the current posture of China, India and Pakistan. (SIPRI)

## Communications, Confidence-Building & Deescalation

**Diplomacy via Track 1 & 2 dialogues between nuclear armed states on use of AI –** The most powerful antidote to escalation or misperceptions is to ensure robust, open dialogue is occurring between states. This is especially important, because as noted by SIPRI, there "remain widespread misconceptions about what AI is and what it can or could do" and it is "difficult for states to assess in a tangible way each other's progress in this area." Beyond bilateral and multilateral diplomatic efforts, formal bodies such as the UN Security Council could also serve as fora for these conversations. One proposal put forth by NTI suggests that the P5 should add the "the technical risks, strategic stability implications, and crisis stability implications of emerging technologies such as AI to their annual agenda." Further, the P5 has recently demonstrated its utility as a group for such discussions with the release of a "*Joint Statement of the Leaders of the Five Nuclear-Weapon States on Preventing Nuclear War and Avoiding Arms Races*" in January 2022, which affirmed that "a nuclear war cannot be won and must never be fought." (SIPRI, NTI, NSCAI, P5)

**Hotline Strengthening –** Beyond ensuring open lines of communication between nuclear armed states, there must be resilient and robust systems with which states can communicate in times of crisis that are compatible with the pace and vulnerabilities of AI systems. When algorithms increase the speed of warfare, and systems can be corrupted by disinformation,

there is a need to have a rapid and trustworthy method of communication to prevent escalation. States should work to develop a “red phone for the 21st century.” One example underway to catalyze such work is the Institute for Security and Technology’s CATALINK initiative. (SIPRI, NTI)

**Development and public disclosure –** disclosing broad policies, strategies and doctrine on how each state intends to use or not use AI in military applications, to the extent that transparency is possible. (SIPRI)

**Design choices for decisions support systems** – In any analyses and decisions support systems connected with NC3, established and declared best practices should be followed. These could include establishing that (1) nuclear launch decisions should not be made on a single source of information and (2) nuclear launch decisions must be verified by human intelligence.

## Strengthening Existing Nuclear Treaties

**Treaty on the Non-Proliferation of Nuclear Weapons** — Strengthen commitments of the NPT’s three pillar structure of non-proliferation, disarmament and promotion of peaceful use of nuclear energy.

**Treaty on the Prohibition of Nuclear Weapons** — Reinforce and strengthen a global norm on the unacceptable dangers of nuclear weapons.

**Bilateral Nuclear Agreements such as New START** — Retain verifiable limits on nuclear forces, and prevent backsliding into a new arms buildup.

**Strengthen the Outer Space Treaty** — The treaty denotes that "states shall not place nuclear weapons or other weapons of mass destruction in orbit or on celestial bodies or station them in outer space in any other manner.”

## Technical Research

- Development of an internationally agreed upon framework for evaluation of the risks of AI in NC3
- Development of robust cybersecurity protocols
- Development of measures to identify disinformation, deep-fakes or manipulated data
- Development of techniques for cryptographically secure, non-proofable, and reliable communication

- Development of transparent, explainable AI decision-support systems
- Development of novel verification methods for software
- Optimization of human-machine interaction to combat automation bias, under-trust of AI systems, and the human out of the loop problem
- Development of standards for machine learning with simulated data
- Development of new, transparent, international standards for robust test, evaluation, validation and verification (TEVV) protocols for AI in NC3
- Research into the development of fail-safes (such as redundant algorithms meant to do the same task but trained with different datasets)