Artificial Intelligence and Nuclear Weapons: Problem Analysis and US Policy Recommendations

14th November 2023

Hamza Chaudhry & Landon Klein
policy@futureoflife.org

The Future of Life Institute (FLI) works to promote the benefits of technology and reduce their associated risks. FLI has become one of the world's leading voices on the governance of artificial intelligence (AI) and created one of the earliest and most influential sets of governance principles, the Asilomar AI Principles. FLI maintains a large network among the world's top AI researchers in academia, civil society, and private industry.

In March of this year, we released an open letter that sparked a global debate on AI safety.

Domain Definition

Since 1945, eight states other than the United States have successfully acquired nuclear weapons: the UK, France, China, Russia, Israel, Pakistan, India, and North Korea. While the possession of nuclear weapons by a handful of states has the potential to create a stable equilibrium through strategic deterrence, the risk of nuclear weapons use on the part of any state actor - and consequent nuclear responses - poses an existential threat to the American public and the international community.

Problem Definition

Developments in artificial intelligence (AI) can produce destabilizing effects on nuclear deterrence, increasing the probability of nuclear weapons use and imperiling international security. Advanced AI systems could enhance nuclear risks through further integration into nuclear command and control procedures, by reducing the deterrence value of nuclear stockpiles through augmentation of Intelligence, Surveillance, and Reconnaissance (ISR), by making nuclear arsenals vulnerable to cyber-attacks and manipulation, and by driving nuclear escalation with AI-generated disinformation.

1. AI Integration into Nuclear Command and Control

As developments in AI have accelerated, some military and civilian defense agencies have considered integrating AI systems into nuclear decision-making frameworks alongside integration into conventional weapons systems with the intention of reducing human error.[1][2] In the United States, this framework is referred to as the nuclear command, control, and communications (NC3) system, which dictates the means through which authority is exercised and operational command and control of nuclear procedures are conducted.

However, a growing body of research has highlighted the potentially destabilizing consequences of integrating AI into NC3.[3] This includes the following threat vectors:

A. **Increased Reliance on Inaccurate Information:** AI systems have already displayed significant inaccuracy and inconsistency across a wide range of domains. As the relevant data that are needed for the training of AI systems for NC3 are extremely sparse - nuclear weapons have only been deployed twice in history, and were deployed in a substantially different nuclear landscape - AI systems are even more likely to exhibit error in these use cases than others. While there has been considerable focus on ensuring that there are 'humans in the loop' (i.e., the final decision is made by a human authority), this may prove to be challenging in practice. If an AI system claims that a nuclear weapon has been launched by an adversary, studies suggest it is unlikely that human agents would oppose this conclusion, regardless of its validity. This problem of 'machine bias' has already been demonstrated in other domains, making the problem

1 Horowitz, M. and Scharre, P. (December, 2019). A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence; Dr. James Johnson on How AI is Transforming Nuclear Deterrence. Nuclear Threat Initiative.

2 Reihner, P. and Wehesner, A. (November, 2019). The real value of AI in Nuclear Command and Control. War on the Rocks.

3 Rautenbach, P. (February, 2023). Keeping humans in the loop is not enough to make AI safe for nuclear weapons. Bulletin of Atomic Scientists.

of ensuring 'meaningful human control' over AI systems incredibly difficult.[4]

B. **Increased Reliance on Unverifiable Information:** At present, it is nearly impossible to determine the exact means by which advanced AI systems reach their conclusions. This is because current means of 'interpretibility' - or understanding why AI systems behave the way they do - lag far behind the state-of-the-art systems themselves. In addition, because modern nuclear launch vehicles (e.g. intercontinental ballistic missiles (ICBMs), submarine-launched ballistic missile (SLBMs)) deliver payloads in a matter of minutes, it is unlikely there would be enough time to independently verify inferences, conclusions, and recommendations or decisions made by AI systems integrated in NC3.

C. **Artificial Escalation and General Loss of Control**: If multiple nuclear powers integrate AI into nuclear decision-making, there is a risk of "artificial escalation." Artificial escalation refers to a type of inadvertent escalation in which adversaries' respective AI systems make calculations based on strategic maneuvers or information originating from other AI systems, rather than from human judgment, creating a positive feedback loop that continuously escalates conflict.[5] Importantly, there is likely to be a dilution of human-control in these situations, as there would be incentives to rely on AI judgements in response to adversary states which are doing the same. For instance, if adversaries are presumed to be making military decisions at machine speeds, to avoid strategic disadvantage, military leaders are likely to yield increasing deference to decision-making and recommendations by advanced AI systems at the expense of meaningful human judgment. This leaves significantly less time for clear-headed communication and consideration, instead motivating first-strike, offensive actions with potentially catastrophic consequences.

2. Expansion of Nuclear Arsenals and Escalation due to developments in Intelligence, Surveillance and Reconnaissance (ISR) capabilities

ISR refers to coordinated acquisition, processing, and dissemination of accurate, relevant, and timely information and intelligence to support military decision-making processes. The belief that other states do not have perfect information about their adversaries' nuclear launch capabilities is essential to maintaining strategic deterrence and reducing insecurity, as it theoretically preserves second strike capabilities in the event of an attack, underscoring mutually-assured destruction. Toward this end, many nuclear powers, including Russia and China, employ mobile missile launchers because they are more difficult to track and target compared to stationary weapons systems. However, both actual and imagined developments in ISR resulting from AI integration increase the perceived threat of detection and preemptive attack on mobile missile launchers and other clandestine military technology. Should a competing nuclear power come to believe that an adversary possesses perfect information regarding the locations of nuclear weapons systems, the possibility that adversaries deploy their nuclear stockpiles rather than risk having them dismantled increases considerably. Such instability is prone to lead to expansion of nuclear arsenals, increased escalation on other fronts, and further risk of nuclear conflict.

4 Baraniuk, S. (October, 2021). Why we place too much trust in machines. BBC News.

5 The short film "Artificial Escalation," released by FLI in July 2023, provides a dramatized account of how this type of escalation can occur. This policy primer delves into mitigation strategies for the risks portrayed in the film.

3. Increased Vulnerability of Nuclear Arsenals and Command Systems to Cyber Attacks

Advancements in artificial intelligence have led to rapid expansion in the capacity for malevolent actors to launch cyberattacks and exploit cyber-vulnerabilities.[6] This includes significantly enhanced capabilities to exploit technical gaps in nuclear security infrastructure (e.g. zero-day vulnerabilities) and to manipulate high-value persons in positions of nuclear command and control (e.g. through deception or blackmail via phishing and spearphishing attacks). NATO allies have pointed out the threat of AI systems being used to attack critical infrastructure, and nuclear arsenals and command and control centers.[7] In addition, if states move toward integrating AI into NC3 systems, such systems would be even more vulnerable to cyberattacks and data poisoning, a practice that entails manipulating the datasets AI systems are trained on to modify their behavior and exploit weaknesses. As data centers and systems are often networked, a cyber-failure could rapidly spread throughout the system, and damage other military command and control systems.

4. Nuclear Escalation and Misperception due to AI-Generated Disinformation

Advanced AI systems have already displayed the capacity to generate vast amounts of compelling disinformation. This disinformation is generated in text using large language models, and via the synthetic construction of fake audiovisual content such as pictures and videos, also known as deep-fakes. Such disinformation is likely to have an outsized negative impact on military confrontation, and in particular on nuclear risk. For instance, if an artificially-engineered piece of audiovisual material is incredibly compelling and signals intended nuclear action, the immediacy of advanced missile technology (see #1B) would not provide sufficient time for vetting the authenticity of the information and may push decision-makers to default to a nuclear response.

Policy Recommendations

In light of the significant risks identified in the previous section, considerable attention from policymakers is necessary to ensure that the safety and security of the American people are not jeopardized. The following policy recommendations represent critical, targeted first steps to mitigating these risks:

1. **Limit use of AI Systems in NC3 and Establish Criteria for 'Meaningful Human Control'**: As recommended by a growing number of experts, the US should prohibit or place extremely stringent constraints the use of AI systems in the highest-risk domains of military decision-making. As discussed, mere human involvement at the tail-end of nuclear decision-making is unlikely to be effective in preventing escalation of nuclear risk from integration of AI systems. Minimizing the use of advanced AI systems where safer alternatives are available, requiring meaningful human control at each step in the decision-making process, and ensuring human understanding of decisionmaking criteria of any systems deployed in NC3,

6 These concerns are discussed in greater detail in "Cybersecurity and Artificial Intelligence: Problem Analysis and US Policy Recommendations".

7 Vasquez, C. (May, 2023), Top US cyber official warns AI may be the 'most powerful weapon of our time.' Cyberscoop; Artificial Intelligence in Digital Warfare: Introducing the Concept of the Cyberteammate. Cyber Defense Review. US Army.

would reduce risks of accidental use and loss of human control, and would also provide crucial signals to geopolitical adversaries that would minimize undue escalation risk.

2. **Require Meaningful Human Control for All Potentially Lethal Conventional Weapon Use:** Escalation to nuclear conflict does not occur solely within the nuclear domain, but rather emerges from broader geopolitical tensions and military maneuvers. Though this brief focuses specifically on the risks at the intersection of AI and NC3, incorporating AI into any military decision-making with major, irreversible consequences increases the risk of artificial escalation and loss of control that could eventually evolve into nuclear conflict. In order to reduce the risk of artificial escalation that could trigger nuclear conflict, the US should require by law that any potentially lethal military decision is subject to meaningful human control, regardless of whether it involves nuclear or conventional weapons systems. While the Department of Defense (DoD) Directive 3000.09 on Autonomy in Weapons Systems presently requires "appropriate levels of human judgment" in the use of force, this could be interpreted to allow for low levels of human judgment in some military operations, and is subject to change depending on DoD leadership. To ensure sound military decision-making that mitigates artificial escalation risk, "meaningful human control" should be codified in statute for any use of potentially-lethal force.
3. **Improve Status Quo Stability by Reducing Nuclear Ambiguities:** The US should formally renounce first strikes - i.e., categorically state that it will not initiate a nuclear conflict - which would help assuage tensions, reduce the risk of escalation due to ambiguities or misunderstanding, and facilitate identification of seemingly inconsistent actions or intelligence that may not be authentic. Finally, the US should improve and expand its military crisis communications network, or 'hotlines', with adversary states, to allow for rapid leadership correspondence in times of crisis.
4. **Lead International Engagement and Standard-Setting:** The US must adopt best practices for integration of AI into military decision-making, up to and potentially including recommending against such integration altogether at critical decision points, to exercise policy leadership on the international stage. In addition, the US should help strengthen the Nuclear Non-Proliferation Treaty and reinforce the norms underpinning the Treaty on the Prohibition of Nuclear Weapons in light of risks posed by AI.
5. **Adopt Stringent Procurement and Contracting Standards for Integration of AI into Military Functions:** Because NC3 is not completely independent of broader military decision-making and the compromise or malfunction of other systems can feed into nuclear escalation, it is vital that stringent standards be established for procuring AI technology for military purposes. This should include rigorous auditing, red-teaming, and stress-testing of systems intended for military use prior to procurement.
6. **Fund Technical Research on AI Risk Management and NC3:** The US should establish a risk management framework for the use of AI in NC3. Research in this regard can take place alongside extensive investigation of robust cybersecurity protocols and measures to identify disinformation. It should also include research into socio-technical mechanisms for mitigating artificial escalation risk (e.g. how to minimize machine bias, how to ensure that military decision-making happens at human speeds) as well as mechanisms for verifying the authenticity of intelligence and other information that could spur disinformation-based

escalation. This would encourage the development of AI decision-support systems that are transparent and explainable, and subject to robust testing, evaluation, validation and verification (TEVV) protocols for specifically developed for AI in NC3. Such research could also reveal innovations in NC3 that do not rely on AI.

Finally, it is vital to set up an architecture for scrutiny and regulation of powerful AI systems more generally, including those developed and released by the private sector for civilian use. The nuclear risks posed by AI systems, such as those emerging from AI-enhanced disinformation and cyberwarfare, cannot be mitigated through policies at the intersection of the AI-nuclear frontier alone. The US must establish an auditing and licensing regime for advanced AI systems deployed in civilian domains that includes evaluation of risk for producing and proliferating widespread misinformation that could escalate geopolitical tensions, and risk of use for cyberattacks that could compromise military command control and decision support systems.