CNA

AI AND NUCLEAR OPERATIONS: IDENTIFYING AND MITIGATING RISKS

Governments around the world have recognized the revolutionary promise of artificial intelligence (AI), with machines completing complex tasks and matching or exceeding human performance. While military applications of AI are inevitable and are already being witnessed, **governments and militaries are relatively silent regarding AI applications to the most dangerous of weapons: nuclear forces**. With nuclear powers modernizing their nuclear forces, it is likely that they will explore nuclear applications as each seeks areas of advantage through AI. How could AI applications—in both nuclear operations and AI-enabled military capabilities more broadly—affect the likelihood of nuclear war or the speed at which conventional conflict could become nuclear?

IDENTIFYING RISKS AND OPPORTUNITIES

Looking to mitigate potential risks, the US State Department's Bureau of Arms Control, Verification, and Compliance (AVC) asked CNA to analyze how AI could impact nuclear risks and what actions could reduce those risks. Building on existing literature, our analysis explored the complex ways that AI could influence nuclear risk—both increasing and reducing risk. We identified mechanisms by which AI-enabled nuclear operations could **increase nuclear risk** or **reduce nuclear risk**, as well as mechanisms by which AI could significantly impact **nuclear risk in an uncertain direction.**

Three challenges that could increase risk stem from **technical characteristics** of AI, the interface between humans and AI (the "**human-machine team**"), and the **ways AI might shape leaders' decisions** about nuclear use in crisis or war:

- **AI technical challenges** include the performance of specific AI systems, complex and unpredictable interactions among AI systems operating in a system of systems, shortcomings in AI training data, poor alignment between AI tools and tasks, and adversarial action against AI systems. These challenges could cause AI to perform unpredictably, increasing nuclear risk.
- **Human interface challenges** include calibrating appropriate human trust in AI, unskilled use of AI by operators, skills degradation, and decision-time compression. All of these challenges degrade the effectiveness of human-machine teaming associated with AI applications.
- **Risks from leader calculus** result from the difficulty of assessing how AI could impact the military balance. Compared to physical capabilities such as warheads and sensors, we lack tangible, measurable indicators for the wide range of potential AI-related capabilities. This uncertainty shapes and complicates leaders' choices, increasing the risk that uncertainty or imperfect information could drive a decision to escalate.

There are also opportunities to use AI to mitigate nuclear risks. We identify four areas for risk mitigation: (1) nuclear weapons surety, (2) survivability and resilience of nuclear forces, (3) leadership decision-time expansion, and (4) crisis and conflict de-escalation. These applications show that applying AI to nuclear operations is not intrinsically risky. Risk is related to the function AI is executing, the specific technical characteristics of the AI application, and the relationship to human operators and decision-makers.

We also describe areas where AI applications could influence risk in an uncertain direction, **either reducing or increasing nuclear risk, depending on the details** of exactly how AI was used, by which actors, and to what ends. The five areas we observed are (1) operation and maintenance of nuclear forces; (2) performance of non-nuclear forces; (3) performance of nuclear forces; (4) analysis, planning, and decision support; and (5) active air and missile defense.

RECOMMENDATIONS

How can nuclear powers avoid bad outcomes and bring about good ones with regard to AI and nuclear operations? We identified three sets of steps that can promote the risk-reducing benefits of AI-enabled nuclear operations. These steps are nested, reflecting the fact that AI applications in the nuclear niche will be shaped by applications in military applications more broadly, as well as in the non-military AI ecosystem. Specifically, we propose the following:

- ***Focused risk mitigation for AI applications in nuclear operations.*** Because of the high stakes and unique characteristics of nuclear operations, some risk mitigation steps should focus specifically on AI applications in nuclear operations. These could include active risk management steps within the US nuclear operations community, engagement with allies and competitors, and agreements among nuclear powers. Best practices would include a commitment to common ethical principles; creation of oversight and governance structures; using wargames and experiments to identify and mitigate risks; reducing risks during training, data curation and algorithm design; and capturing and acting upon lessons learned.
- ***Applied efforts on risk mitigation of AI applications in military operations.*** The US military, like many militaries around the world, is applying AI to many functions involving conventional warfare. These general military applications will share many of the same challenges as those in nuclear operations. Military and government leaders responsible for nuclear operations should work with the US military to reduce risks from military applications of AI overall. The US government can also work with foreign allies and partners to this end through technical cooperation. Broader international agreements and discussions of AI safety and ethics may also help with nuclear risk reduction.
- ***Basic research and practical solutions for fundamental sources of AI-related risks.*** Given the relative newness of modern AI techniques and the focus on commercial applications versus fundamental understanding and safety, there are many aspects of AI risks that are still not well understood. Widely acknowledged challenges include human-machine teaming and trust, AI machine-machine interactions, and validation of appropriate areas of AI applications. The US government can work with other governments, industry, and academia to better understand these risks and to seek collective solutions to mitigate them.

CONCLUSIONS

Our detailed characterization of these mechanisms and their potential consequences provides a broader and deeper exploration of the complex relationship between AI and nuclear risk than can be found in the existing literature. Based on the interactions among such complex identified factors, we could not make an absolute conclusion regarding whether the net effect of AI-enabled nuclear operations will be positive or negative. The details of what countries choose to do in the AI-nuclear space and exactly how they do it will matter a great deal—and it is by no means clear today what path each country will take. However, the steps provided above can help guide nuclear powers, and militaries overall, to have AI applications reduce risks associated with nuclear operations.

ABOUT CNA

CNA is a nonprofit research and analysis organization dedicated to the safety and security of the nation. It operates the Center for Naval Analyses — the federally funded research and development center (FFRDC) of the Department of the Navy — as well as the Institute for Public Research. CNA develops actionable solutions to complex problems of national importance.

This is a short summary of CNA's research on this topic.
For more information, contact:

Dr. Larry Lewis | 703.725.3633 | lewisl@cna.org or Dr. Tim McDonnell | 703.824.2339 | mcdonnellt@cna.org

IMM-2023-U-034812-FINAL