Real-Life War Games: Fully Autonomous Nuclear Weapons Systems

"Artificial intelligence is the future, not only for Russia but for all of humankind. Whoever becomes the leader in this sphere will become the ruler of the world" (Putin, personal communication, September 1, 2017).

Even in its infancy, artificial intelligence has revolutionized warfare. Great powers including China, the United States and Russia have automated their weaponry and nuclear weapons systems to include "narrow-AI," which refers to limited, AI-assisted tasks such as autonomous retargeting. Narrow-AI systems defer to a human commander before firing and can be overridden. However, the trend towards automation emboldens the risk of future development of fully autonomous nuclear weapons systems. Autonomous systems "select and engage targets without further intervention by a human operator," meaning that there is no human making the final decision to attack a target (Saylor, 2021). With the implementation of AI into nuclear command and control (NC3), nuclear weapons systems will become fully autonomous. AI will then control the initiation, trajectory and outcome of nuclear war. The potential development of autonomous nuclear weapons systems is an urgent prospect that ought to be addressed by the international community.

AI has inherent fallibility in war. Machine learning depends on past neural networks developed through pattern recognition within controlled environments. Absent perfect simulations, AI may create false positives when encountering new data. In contrast to humans, AI cannot recognize the function of an object by appearance, and therefore cannot differentiate between vehicles that carry nuclear missiles versus those that do not. Due to inherent vulnerabilities, AI is a prime target of spoofing and hacking (Loss, 2019). Human interaction with AI is especially complicated. During the Cold War, the Soviets trained their computer software with information which reflected their pre-existing beliefs about the US, resulting in output that reinforced their suspicions. This phenomenon leads to "automation bias," as leaders are more likely to trust AI with built-in biases (Horowitz, Sharre & Velez-Green, 2019). Rife with technological shortcomings, AI is devastating in war.

Most significantly, AI lacks intentionality and political context, which may undermine deterrence. In the past 70 years, nations practiced deterrence because they feared mutually assured destruction, thus refraining from striking first because retaliatory strikes could destroy their own nation (Hymans, 2015). In *The Precipice*, Toby Ord cites several near-crises which were avoided due to human override of narrow-AI systems. In 1983, Soviet officer Lt. Col. Petrov chose to disregard warnings delivered by a narrow-AI system that US nuclear missiles were approaching. The warnings turned out to be a false alarm. AI does not understand false alarms, and it is difficult for humans to correct or stop malfunctioning AI (Spindel, 2020). In this crisis, humanity prevented a nuclear war that AI would have initiated. If humans are ultimately not in control of weapons launch, autonomous systems are more likely to go to war. War will proliferate at machine speed, and escalate absent hesitation. Global leaders may not be able to de-escalate in time, and humanity could be at the mercy of the weapons they created.

Autonomous nuclear weapons systems could be impending (Futter, 2020). If nations suspect that competing nations will automate their armaments, they will be incentivized to automate their own, initiating a cycle of arms racing driven by perception (Boulanin, 2018). Russia is purportedly developing an autonomous nuclear torpedo, Poseidon, although these claims are unconfirmed. The US Department Of Defense disclosed plans to highly automate NC3 systems to reduce the role of humans in weapons launch in 2021. The Pentagon's 2020 report emphasized the development of autonomous systems, as machine-speed command and control will play a decisive role in future warfare (Klare, 2020). Nations including China are exceptionally likely to pursue autonomy due to insecurities about their second-strike capabilities, or ability to retaliate once an opposing nation strikes. China's NC3 system includes partially automated command and control to reduce human interaction (Cunningham, 2019). Military AI progression for China is their strategic advantage compared to the US (Allen, 2019). If China perceives they could challenge US military primacy with AI weaponization, China's interest in autonomous nuclear weapons may accelerate (Denmark & Talmadge, 2021). In an arms race between great powers, each may install autonomous NC3 without communicating the nature or status of the systems (Klare, 2020). Absent understanding of how the weapons are used and will interact, actions will be inherently unpredictable and nations could misinterpret signals. Considering these strategic problems, international actors must intervene.

Global leaders should engage the international community in cooperative efforts, focusing on a common moral foundation and obligation to avoiding war. International institutions will be central to discussions of existential risk mitigation relating to nuclear war. Moral obligations could be presented in a constitution for humanity articulating the paramount need to safeguard humanity's future (Ord, 2020). An international governing body including global leaders and AI experts could monitor developments, facilitate cooperation, and set limits. Autonomous nuclear weapons systems could potentially be banned, similar to how experts have called for bans on lethal autonomous non-nuclear weapons (Gubrud 2014). However, a direct ban may be premature (Ord, 2020). Instead, international actors should require "meaningful human control" in war, with humans making the final decision to use force. An autonomous weapons convention should frame the agreement through strong, intuitive moral principles. The agreement would include relevant definitions and exceptions, and support a treaty implementing organization (Gubrud 2014). Such mitigatory efforts are imperfect, and their shortcomings must be acknowledged.

A major associated issue to be addressed is circumvention of the ban by belligerent nations. Compliance could be enforced through transparency measures and robust compliance checks. Effective verification would require inspecting both software and hardware of weapons systems. Even considering these measures, inherent risks remain. It is impossible to determine if a weapons system is autonomous by appearance alone, and notoriously difficult to read software. Software may additionally be encrypted (Gubrud 2014). Trust between nations must be built through other means.

Confidence-building measures could help manage remaining risks. Low cost changes improve securitization and international coordination to lower the risk of nuclear war (Ord, 2020). Nations should increase transparency about intended use of weapons systems, and develop lines of communication similar to the hotline between the US and the Soviets amid the Cold War (Spindel, 2020). International institutions should develop norms including prohibiting the targeting of autonomous nuclear systems and disincentivizing the weaponization of AI as a best practice to maintain cybersecurity (Avin & Amadae, 2020). Norms insert diplomacy back into warfare to reduce mistrust or miscommunication between nations.

Nations should continue to work towards eliminating or decreasing their nuclear arsenals. Restarting arms reduction agreements such as the Intermediate-Range Nuclear Forces Treaty could reduce nuclear risk. Nations including the US could unilaterally remove their ICBMs from hair trigger alert to decrease the speed of war and thus the likelihood of accidents (Ord, 2020). Partial denuclearization may be an idealized goal, but incremental efforts to reduce nuclear risk may be more politically viable. Exercising caution with upcoming technologies will be a central piece of the puzzle. With the potential development of autonomous nuclear weapons systems, pre-emptive discussion and negotiation must facilitate peace before a crisis arises.

References

Allen, G. (2019, December). *Chinese strategic intentions: A deep dive into China's ... - nsiteam.com*. NSI. Retrieved May 18, 2022, from https://nsiteam.com/social/wp-content/uploads/2019/10/SMA-Chinese-Strategic-Intentions-White-Paper-FINAL-01-Nov-2.pdf

Avin, S., & Amadae, S. (2020, April 1). Autonomy and machine learning at the interface of nuclear weapons, computers and people. In Boulanin, Vincent. Stockholm International Peace Research Institute, The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk. [Book chapter]. https://doi.org/10.17863/CAM.44758

Cunningham, F. (2019, July 19). *Nuclear command, control, and communications systems of the People's Republic of China: Nautilus Institute for Security and Sustainability*. Nautilus Institute for Security and Sustainability. Retrieved May 17, 2022, from https://nautilus.org/napsnet/napsnet-special-reports/nuclear-command-control-and-communications-systems-of-the-peoples-republic-of-china/#_ftnref50

Saylor, K. (2020, December 1). *Defense primer: U.S. policy on Lethal Autonomous Weapon Systems*. Congressional Research Service. Retrieved May 18, 2022, from https://sgp.fas.org/crs/natsec/IF11150.pdf

Denmark, A. M., & Talmadge, C. (2022, February 19). *Why China wants more and Better Nukes*. Foreign Affairs. Retrieved May 17, 2022, from https://www.foreignaffairs.com/articles/china/2021-11-19/why-china-wants-more-and-better-nukes

Futter, A. (2020, October 15). *Artificial Intelligence, Autonomy and Nuclear Stability: Towards a More Complex Nuclear Future*. Valdai Club. Retrieved May 17, 2022, from https://valdaiclub.com/a/highlights/artificial-intelligence-autonomy-and-nuclear-stability/

Gubrud, M. (2014, January 4). *Can an autonomous weapons ban be verified?* ICRAC. Retrieved May 17, 2022, from https://www.icrac.net/can-an-autonomous-weapons-ban-be-verified/

Horowitz, M.C., Scharre, P., & Velez-Green, A. (2019). A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence. ArXiv, abs/1912.05291.

Hymans, J. (2018, June 28). *The Psychology of Nuclear Restraint*. Bulletin of the Atomic Scientists. Retrieved May 17, 2022, from https://thebulletin.org/2015/10/the-psychology-of-nuclear-restraint/

Klare, M. T. (2020, April). *Arms control Today*. 'Skynet' Revisited: The Dangerous Allure of Nuclear Command Automation | Arms Control Association. Retrieved May 17, 2022, from https://www.armscontrol.org/act/2020-04/features/skynet-revisited-dangerous-allure-nuclear-command-automation

Loss, R. (2019, September 19). *Will artificial intelligence imperil nuclear deterrence?* War on the Rocks. Retrieved May 17, 2022, from https://warontherocks.com/2019/09/will-artificial-intelligence-imperil-nuclear-deterrence/

Ord, T. (2021). The Precipice. Hachette Books.

Vincent, B. (2018, December 7). *AI & Global Governance: AI and nuclear weapons - promise and perils of AI for Nuclear Stability*. United Nations University Centre for Policy Research. Retrieved May 17, 2022, from https://cpr.unu.edu/ai-global-governance-ai-and-nuclear-weapons-promise-and-perils-of-ai-for-nuclear-stability.html