ML/AI and weapons of mass destruction

We are now in a situation where historically neutral countries Sweden and Finland have submitted their applications to join NATO, two relatively big European nations are in open conflict for the first time since the end of the Second World War, and there is potential that one side may feel increased incentives to resort to nuclear use if conventional means of war fail. We should ask ourselves if the mutually assured destruction (MAD) and assured retaliation that caused relative peace between nuclear-armed states is still assured or if some states that possess nuclear arsenal may feel threatened. In this work, I describe current known automation and future possible impacts of AI/ML on use in nuclear command and control, early warning systems, intelligence, surveillance, and reconnaissance, and also in the nuclear launch platforms and delivery vehicles.

Historical stability

As is described by Geist et al. mutually assured destruction was (is) a condition that both USSR and USA hoped to escape if possible. In the Reagan era, the USA for example tried to create a missile defense while the USSR developed a civil defense program which would assure survival of at least part of the population. MAD did deter a Soviet preemptive strike but did not deter the Soviet Union from using its conventional superiority to attack western Europe. The USA therefore developed a doctrine of assured retaliation that assured appropriate response for any enemy provocation. As the nuclear arsenal of the U.S. grew to reassure the European allies of the deterrence, the Soviets believed that Americans might develop first-strike capability to attack the Soviets without fear of a devastating retaliation. There was therefore a need for reassurance that adversaries will not be attacked if they refrain from deterred behavior.

The stability therefore relies on the credibility of deterrence, assurance, and reassurance, and there are many ways how AI could help maintain this credibility and on the other hand many ways how AI can destabilize the current relatively peaceful era.

Current state of automation

In order for us to objectively evaluate the possible impact of future AI/ML on stability we will describe automation bias and automation that is or was already implemented in the current hierarchy of command and control as well as in early warning and missile targeting.

Important factor of how beneficial automation will be is how much does the operator trust the machine. There can be a trust gap when for example ground military does not trust unmanned drones and prefers pilots who in their opinion perform more effectively because they have more “skin in the game” even though there is no evidence of their effectiveness. On the other hand, once humans believe in the effectiveness of the automation, they are

more willing to obey the judgment of the machine even when there is evidence that the machine may be incorrect in some situation. This is called automation bias.

Horowitz et al. and Geist et al. describe many examples of current and historical automation. During the Cold War, there was a near-accident in 1983 due to a false alarm by the Soviet Oko automated early warning system which registered five U.S. ICBM launches. Fortunately, the watch officer did not succumb to an automation bias and did not listen to the alert system which reported the missile strike with the highest confidence.

Another automation system in the Cold War era called VRYAN was a computer program supposedly developed by KGB and it would track correlation of forces between U.S. and USSR based on data from intelligence officers and would notify leaders when a preemptive nuclear strike would be needed in order to prevent the U.S from decisive military superiority. This program was flawed in the sense that the data provided for the VRYAN were prejudiced so that they would conform with the leadership's view. This led to a sort of confirmation bias or a feedback loop that reinforced Soviet's fears of U.S. first-strike superiority. The VRYAN was tested in the late 1970s and in 1983 during NATO's annual exercise allegedly led to Moscow placing forces on higher readiness out of fear that the exercise was the start of U.S. nuclear preemption.

In 2014, Russia founded the National Defense Control Center (NDCC) which functions as information fusion in support of conventional and nuclear operations. The Russian government is simultaneously investing heavily in AI, among other reasons, to better process the large amount of data the NDCC has to process. These investments are occuring in the time when Russia possibly failed to achieve its objectives through conventional war in Ukraine.

Another possible (Horowitz et al. mentioned that the source of this information is interview with former Soviet officer) automation from the Cold War era is the Soviet dead hand system called "Perimeter". A dead hand system can launch a nuclear counterattack even in case when leadership is wiped out. The "Perimeter" specifically was allegedly a network of sensors that could detect a nuclear explosion on Soviet territory and after 15 minutes to an hour without a halt order from leaders, it would transfer nuclear launch authority to duty officers in bunkers. They would then launch rockets over Soviet territory which would beam down launch codes to other missile silos. But this is only a speculation because Soviets did not communicate the possession of the dead hand system to the U.S. leaders. This would be counterproductive as the U.S. would not know about certain retaliation from the Soviet side.

The Russian missiles after the Cold War era would in case of launch without a flight plan allegedly automatically revert to their wartime targets in the U.S.. The same automation is supposedly used in U.S. missiles which can be retargeted in 10 seconds.

Today there are not only Russia and the U.S. but also China which is heavily investing in decision support systems and AI in general. Horowitz et al. refer to Lora Saalman who says that China fears "bolt-from-the-blue" preemptive attack on its nuclear forces and therefore prioritize avoiding "false negative" over "false positives". As China will rely heavily on AI, the Chinese officials may be susceptible to automation bias.

As is shown in the cases of near-accident with Oko and VRYAN escalation, even the Cold War era automation led to risk of automation bias. Question is what will happen in future when AI is expected to become more widely used in decision support systems?

Future state of automation

Zero data problem

As there was never a preemptive nuclear strike there is zero data on which we could train our machine learning models to recognize the strike and even more we don't even know what would be a reliable indication of such a preemptive strike. Even though the AI-based decision support system may work perfectly during peacetime, strengthening the automation bias of the army generals, we don't know how the support system will behave in the most critical moment of a crisis.

There may be for example some indirect positive indication that large-scale military mobilization is underway, but we can't see whether decision-makers want to launch nuclear first-strike. In the worst case, the adversary nation which seems ready to launch a preemptive attack might in fact be just preparing itself for defense in response to our higher alert status. This feedback loop of incremental escalation could lead to a mutual nuclear destruction.

Machine learning based decision support systems in conclusion lack the data on which to train and therefore can at best recognize out of the norm situations. If we use more simple alert systems with more transparent reasoning that leaves the analysis part to humans, the chance of automation bias is reduced but at the cost of increased delay between the first alert and final decision, and furthermore, the homan analyst can have his own biases.

Nuclear Launch Platforms and Delivery Vehicles

The nuclear launch platforms can be for example submarines, aircrafts, missile launch facilities, or transporter erector launchers. These systems launch the delivery vehicles which can be for example bomb, missile, or torpedo that carries the nuclear warhead to target.

Nuclear launch platform (NLP) automation was historically very limited (if there was any automation). There is currently a B-21 bomber in development which will be capable of unmanned flight and could carry nuclear weaponry (its size and shape is similiar to the B-2 bomber capable of carying nuclear weaponry). But the trust gap of U.S. officials is too high and the other means of nuclear superiority are too good to risk sending unmanned B-21 simultaneously with the nuclear weaponry as with increased complexity of automation comes risk of hacking, bugs, and other possible vulnerabilities that could be avoided by preserving humans in the loop. However, other nations with less resources that do not have military superiority could calculate risk/benefit ratio differently.

There are a lot of benefits of for example nuclear-armed UAVs. First, they have longer possible endurance, which decreases the possibility of a successful first strike disarming. Second, the UAVs already in air can deliver nuclear weapons faster since they can be closer to a potential target with less frequent maintenance landings. Third, the UAVs can be cheaper and with greater range compared to the manned counterparts.

One country for which the autonomous NLP/Delivery Vehicle could be beneficial is Russia as on March 1, 2018, Russian President Vladimir Putin confirmed the existence of Status-6 which is characterized by U.S. Dept. of Defense as an intercontinental, nuclear-armed, nuclear-powered, undersea autonomous torpedo. Horowitz et al. also mention that Status-6 could use AI to evade enemy anti-submarine warfare forces (source: Vladimir Tuchkov). This torpedo would be limited by the fact that it cannot hit any ground target and therefore poses a danger only to coastal targets, ships, aircraft carriers, and probably to submarines. The benefits to Russia's deterrence tactics is marginal and can mainly cause escalation risks. If the Status-6 cannot be recalled and the transit time is approximately over one day from Denmark Strait and over two days from Barents Sea to New York, then the Russian government seriously limits its ability to negotiate war termination for the crucial two days of conflict.

In the case of unmanned launch platforms, the decision to start a nuclear war is delegated through onboard automation to a human operator who does not have to have any "skin in the game". Unmanned vehicles could cause other risks such as the chance of malfunction or possibility of hacking. In case of communication failure the UAV could be automated to return home, but it cannot be guaranteed that there wouldn't also be simultaneous loss of other sensor data (for example position data) and the aircraft could land in the hands of adversaries along with the sensitive technology and nuclear warhead. Given the importance of the UAV technology and the nuclear armament, the reward for successful hacking is large. In case of the success of third party hacking, the potential is enormous. The state which gets the UAV under control could attack another country and cause escalation between the two targeted countries. Even in cases where the external communication is lowered to a minimum, there is still a chance for successful hacking.

The Stuxnet attack (as described by Kushner D.) can be seen as an example of successful hacking even without any external connection to the Internet. The malware infected the Iranian uranium fuel enrichment plant through human error and used four zero day vulnerabilities in order to slowly wear and tear the enrichment centrifuges by increasing and decreasing its rotational speed. The two main suspects of this attack are Israeli and U.S. governments (Kushner D.). This attack of possibly two governments on one that posses enrichment facilities for production of (high concentration of) U235 that is key part in production of nuclear warhead (with enough centrifuges) shows that with high enough motivation even critical infrastructure and possibly highly automated nuclear launch platforms with average cybersecurity protection can be hacked.

The nuclear launch platforms may therefore need to balance between autonomy and probability of hacking vulnerability. With full autonomy there is lower risk of intrusion, but the system would not be recallable. If there is possibility of communication with the autonomous delivery vehicle (/launch platform) then the communication introduces new ways of intrusion, the system is more prone to communication and technical malfunctions, unanticipated

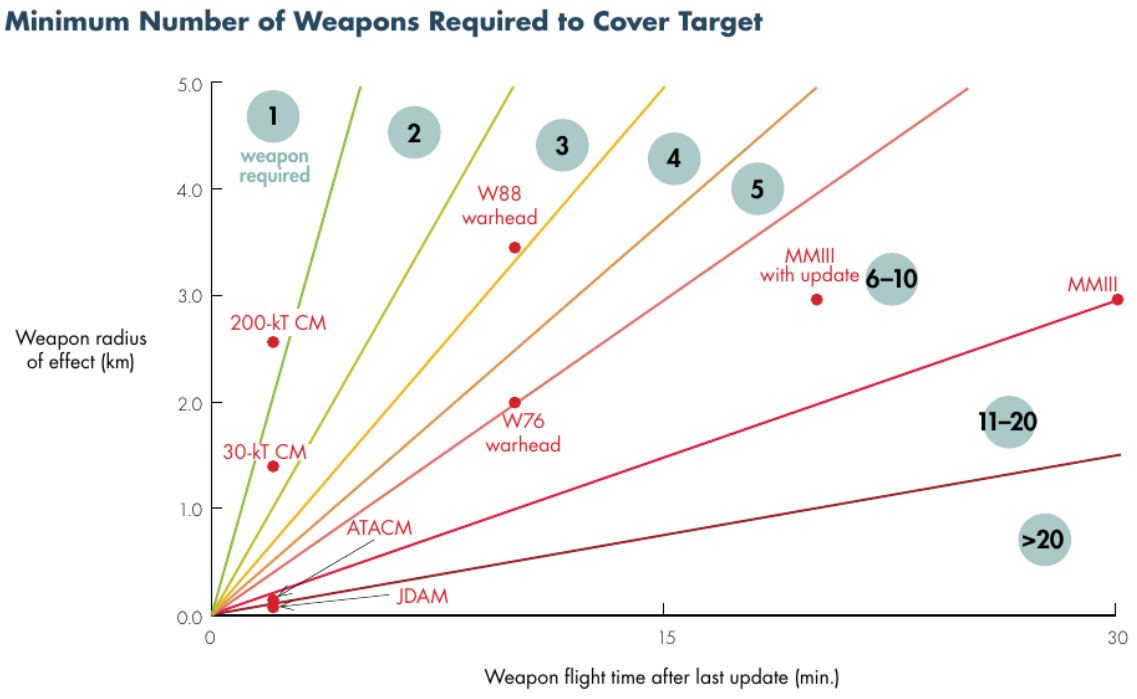
environmental interaction, and possible interference of communication increasing the risk of accidental escalation.

Conventional autonomous systems

States are more likely to accept increased autonomy in conventional systems. Some of these conventional applications can be used in operations concerning nuclear arsenal in direct or indirect ways.

Autonomous systems for launch platforms tracking

Horowitz et al. argue that conventional autonomous systems could be used for real time tracking of the whole adversary's nuclear arsenal but not for successful fast and accurate first-strike operations. They argue that if one country achieved full transparency of adversaries all Transporter Erector Launchers (TEL) it would be at cost of decreased confidence in the survivability of the adversary mobile ICBMs and therefore undermining its deterrence and destabilizing the MAD state. This conventional use of AI in intelligence, surveillance and reconnaissance could have strategic consequences if it would be overused. But TELs are one of the easier targets to track. In order to be successful in the first-strike operation and achieve a state where the adversary cannot retaliate, we would need to target and track all nuclear launch platforms including for example nuclear-powered ballistic submarines (SSBN) with nuclear warheads. In case of tracking all NLPs we would need many autonomous UAVs and again, we encounter the same kinds of problems as with the autonomous NLPs like possibility of malfunction, hacking, communication problems and interference, and additionally inaccuracy of collected and processed information. And even when it would be possible to accurately target all the adversary NLPs, there would still be the problem of delivering sufficiently destructive and accurate weapons in time before it is able



This figure shows the number of warheads of various types that would be required to destroy a mobile target with a weapon radius of effect between 0 and 5 kilometers. Despite their huge "kill radius" measuring kilometers in diameter, multiple thermonuclear warheads delivered by ballistic missiles would be required to have a high assurance of destroying a missile launcher. ATACM = Army Tactical Missile System; JDAM = Joint Direct Attack Munition; kT = kiloton; MMIII = Minuteman III.

to strike back or conceal itself. Even one missed nuclear launcher would mean the possibility of destruction of a whole large city. As can be seen on the figure above in order to achieve high assurance of destroying a NLP we would in some cases need multiple nuclear warheads to destroy even one NLP. This possibility of launcher survival could be enough to deter the attempt for first-strike but even the tracking part could change the behavior of adversaries in a crisis because of their perceived vulnerability of their own arsenal.

Let us now imagine that tracking of all ground and aerial NLPs is possible and let us focus on the submarine NLPs. In order to make the ocean “transparent” we would need a huge number of autonomous reconnaissance vehicles and many sensor fields. Even in the case of Chinese and British submarines which must pass through chokepoints, we would still need a high number of sensors and even without submarines China is still able to cover U.S. targets with its land-based ICBM launchers. And if we compare the armies of China and the U.S. in my opinion even the surveillance of these chokepoints would mean escalation to conventional war or even nuclear retaliation. Even if we are able to classify and localize one SSBN, it could still conduct countermeasures to degrade the ability of the sensor to maintain the tracking. Let alone we need to localize and simultaneously destroy all the adversary SSBN’s to avoid retaliation. Even if these scenarios are not realistic or even impossible in near future, adversaries only need to perceive themselves as vulnerable to advanced technology to escalate the situation.

Other use cases of AI in counter-nuclear operations are for example enhanced missile defenses, improvement of accuracy of counterforce strike options, improved efficacy of cyber-attacks against enemy nuclear command-and-control systems.

The Security of Time

As was said at the beginning, there were already some attempts to automate the delegation process by means of a dead hand system in case of destruction of command centers. This process could be further developed in cases of countries that could feel threatened by a faster pace of conventional wars that could be caused by further automation. The faster pace could also force some vulnerable countries to use preemptive nuclear strike sonner before facing a conventional defeat. In the far distant future, for example a country that posses autonomous plane might be able to avoid air defense threats, easily defeat adversaries in air-to-air engagement and complete its missions much faster. This could even result in so-called “battlefield singularity,” in which battles would be so fast that they would be faster than human decision-making and therefore humans would lose all control above the situation. The fear of losing quickly could be an incentive to use preemptive or faster nuclear strikes.

On the other side, automation could make nuclear escalation less likely. Many countries deploy large networks of sensors which produce large amounts of data. The newly emerged machine learning and other AI capabilities could declutter the information and help the decision-makers make far better sense of the battlespace.

Conventional superiority

Horowitz et al. argue that when a country achieves conventional superiority by means of autonomous robotic systems, there are countries like Russia or Pakistan which indicate in their doctrine that they are willing to use limited nuclear strikes to end a conventional war that they are losing on favorable terms. The robotics and autonomous systems also possibly narrow the gap between nuclear powers. As the current key driver of AI and robotics is the commercial sector, AI could balance the differences between superpowers' level of development.

Conclusion

I will now share my opinion on this topic. Many of the mentioned possible uses of automation will probably be impossible even in the next 50 years. My main takeaway from this is that it is impossible to have first-strike capability without possibility of nuclear retaliation when the adversary possesses SSBN with nuclear warheads. The countries which are more vulnerable to first-strike (and also vulnerable in a general sense) will be more likely to use automation and autonomous/unmanned vehicles. Even though Horowitz et al. saw heavy increase in Russian spending on AI it does not look like Russia has much of an advantage in the current conventional war against Ukraine. I am not a security expert and cannot see through propaganda so this is only a speculation. Only “advanced” drones that I have seen in use by the Russian army were old plastic models with canisters and commercial cameras in them, so projects like Status-6 are only fear mongering and too advanced technology. Russia could therefore feel very vulnerable and there is a probability of some sort of escalation. I also think that only the U.S. and China are possible superpowers that could one day challenge each other with autonomous systems but it is still a very distant and unpredictable future.

References

Geist, E., Lohn, A., J. (2018). How Might Artificial Intelligence Affect the Risk of Nuclear War. https://www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE296/RAND_PE296.pdf

Horowitz, C. M., Scharre, P., Velez-Green, A. (2019). A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence. https://arxiv.org/ftp/arxiv/papers/1912/1912.05291.pdf

Kushner D. (2013). The Real Story of Stuxnet: How Kaspersky Lab tracked down the malware that stymied Iran’s nuclear-fuel enrichment program, https://courses.cs.duke.edu/spring20/compsci342/netid/readings/cyber/stuxnet-ieee-spectrum.pdf

Mizokami K. (2016). Experts: North Korea May Be Developing a Dirty Bomb Drone, http://www.popularmechanics.com/military/weapons/a24525/north-korea-dirty-bomb-drone/.

Figure reference

Figure 1:
Geist, E., Lohn, A., J. (2018). How Might Artificial Intelligence Affect the Risk of Nuclear War. https://www.rand.org/content/dam/rand/pubs/perspectives/PE200/PE296/RAND_PE296.pdf