AI Nuclear Winter or AI That Saves Humanity? AI and Nuclear Deterrence

Nobumasa Akiyama

Contents

Abstract

Nuclear deterrence is an integral aspect of the current security architecture and the question has arisen whether adoption of AI will enhance the stability of this architecture or weaken it. The stakes are very high. Stable deterrence depends on a complex web of risk perceptions. All sorts of distortions and errors are possible, especially in moments of crisis. AI might contribute toward reinforcing the rationality of decision-making under these conditions (easily affected by the emotional disturbances and fallacious inferences to which human beings are prone), thereby preventing an accidental launch or unintended escalation. Conversely, judgments about what does or does not suit the "national interest" are not well suited to AI (at least in its current state of development). A purely logical reasoning process based on the wrong values could have disastrous consequences, which would clearly be the case if an AI-based machine were allowed to make the launch decision (this virtually all experts would emphatically exclude), but grave problems could similarly arise if a human actor relied too heavily on AI input.

Keywords

Deterrence · Escalation · Nuclear · Stability · Weapons

Introduction

Technological innovation often brings about paradigm shifts in various dimensions of our life. Artificial Intelligence (AI)

N. Akiyama (✉)
Graduate School of Law /School of International and Public Policy
Hitotsubashi University, Tokyo, Japan
e-mail: n.akiyama@r.hit-u.ac.jp

J. von Braun et al. (eds.), *Robotics, AI, and Humanity*, https://doi.org/10.1007/978-3-030-54173-6_13

certainly has great potential to fundamentally transforming various dimensions of our life, for better or worse (Kissinger 2018).

The military and security domains are no exception. The trajectory of AI development, together with that of complementary information technology and others, will have a large effect on security issues. AI could make any type of weapons and military system smarter, and it may make warfighting more efficient and effective. The potential of military applications of AI is enormous, and some have already materialized.

Meanwhile, we should be aware of limitations that AI systems may have. The autonomy in weaponry systems poses serious and difficult questions regarding the legal and ethical implications, the evolution and changes of military doctrines and strategy as well as the risks of misguiding decision-making and actions, and the balance of power among nations. The discussion on how to regulate the military application of AI confronts us with significant political challenges.

Among many security-related questions, how the possible convergence of AI and nuclear strategy/deterrence will transform our security environment is the most important one. Nuclear weapons were a game changer in altering the nature of war and security strategy (Jervis 1989). Now AI may effect another game change, when it is integrated into nuclear weapon systems. As nuclear weapons keep posing existential threats to the human being, and the stable, safe management of nuclear arsenal and threat and risk reduction are essential for the survival of human beings, it is natural to ask what the growth of AI will mean for nuclear deterrence, nuclear threat reduction, and nuclear disarmament. Would AI reinforce nuclear deterrence and strategic stability, or undermine them? Would AI promote nuclear disarmament and nonproliferation? To begin with, how could AI be utilized in nuclear strategy and deterrence?

According to a study by RAND Corporation, there are three main positions on the potential impact of AI on nuclear stability. "Complacents" believe that AI technology would never reach the level of sophistication to handle the complicated challenges of nuclear war, and therefore AI's impact is negligible. "Alarmists" believe that AI must never be allowed into nuclear decision-making because of its unpredictable trajectory of self-improvement. "Subversionists" believe that the impact will be mainly driven by an adversary's ability to alter, mislead, divert, or trick AI. This can be done by replacing data with erroneous sample or false precedent, or by more subtly manipulating inputs after AI is fully trained (Geist and Andrew 2019). This categorizing effort suggests that even experts cannot reach a consensus about the potential trajectory of AI, and therefore it is difficult to assess the impact of the technology on nuclear stability and deterrence.

At most, what we can do under such a circumstance where forecasts of the consequences of AI development for the international security environment are necessarily tentative is to present the points of concerns which may arise from various scenarios. So, what factors will shape the future of nuclear deterrence and stability in the era of rise of AI? This paper outlines the risks associated with the possible application of this evolving technology into nuclear security, and the possibilities of contribution to the reduction of nuclear risks.

First, it shows how AI could function as a decision-support system in nuclear strategy. Second, I indicate how AI could support nuclear operations with improvements in intelligence, surveillance, reconnaissance (ISR), and targeting. The virtue of AI can be best demonstrated in a constellation of new technologies, and ISR is the most critical field where such constellation happens. In this connection, although it is important to analyze how AI application to conventional weapons such as Lethal Autonomous Weapon Systems (LAWS) or swarming unmanned aerial vehicles (UAV) bring about changes in the role of nuclear weapons in military strategy, the discussion will be left for another occasion. And third, I show how AI will affect the discourse on ethics and accountability in nuclear use and deterrence.

AI in Supporting Nuclear Decision-Making

Essence of Nuclear Deterrence and the Role of AI

Goals to achieve by nuclear weapons are complex. A fundamental/original purpose of the weapon is to win a warfighting. With its huge destructive power, nuclear weapon could be an effective tool to lead a country to victory. Meanwhile, exactly because of this feature, there are high hurdles for the actual use of nuclear weapons as it might cause humanitarian catastrophe not only in the adversary's population, but also in its own population when retaliated. Thus, the advent of nuclear weapons has changed the nature of war, and is employed to achieve certain political goals without detonating one in warfighting, or through nuclear deterrence.

Deterrence is the use of retaliatory threats to dissuade an adversary from attacking oneself or one's allies. Deterrence is not the only goal nor the winning a war. To succeed, deterrence should be multidimensional. Nuclear weapons can also be used for compellence—coercing the enemy into doing something against its will. In the meantime, the coercive elements are not the only components of a strategic relationship between potential adversaries. Both sides must be convinced to certain degree that they will not be attacked

as long as they would not provoke beyond some level of confrontation that is being deterred (Schelling 1966).

In particular, between major nuclear powers in an adversarial relationship, the maintenance of strategic stability is sought in order to avoid armed conflict, which could be escalated into a catastrophic nuclear war if they fail to manage the escalation. Strategic stability is understood as "a state of affairs in which countries are confident that their adversaries would not be able to undermine their nuclear deterrent capability" and that state would be maintained with various means including nuclear, conventional, cyber, or other unconventional means (Podvig 2012).

Another important feature of nuclear deterrence, which requires thorough consideration is that effective deterrence needs to complicate adversary's strategic calculus, but to the extent that it would not challenge the *status quo*.

Beyond military capabilities, postures, and pressure, effective deterrence requires, to a greater extent, the sophisticated skill of political communication and crisis management. To achieve the objectives of deterrence, it is necessary to make the adversary lose confidence in the success or victory of its strategy and, at the same time, to make sure that the adversary would be convinced that its core national interests would be preserved if the situation would not be inflicted into the use of force. Although it is essential for successful deterrence to demonstrate strong resolve and sending clear message that action corresponding to such a resolve would be taken, and it will be ready to give damage imposing unbearable cost on the adversary, it would require compromise and carefully avoid total humiliation of an adversary, which may drive the adversary into desperate action.

So where does AI fit in? As referred above, with machine learning, AI optimizes its performance to achieve a goal provided. Systems using these techniques can, in principle if not in practice, recursively improve their ability to successfully complete pattern recognition or matching tasks based on sets of data (which usually need to be carefully curated by humans first) (Heath 2018).

Meanwhile, current AI systems may still have limitations in performing when they are operated outside the context for which they are designed to work, and transferring their learning from one goal to another. This feature suggests that AI at the current level of its capabilities would not function well. In a complex social and political environment where the final goals of action and the choice by decision makers are adaptable to emerging and evolving conditions surrounding decision makers, it is unlikely that, under the current level of technical competence and maturity of discussion on ethical concerns, AI alone will/can make a decision to use nuclear weapons.

AI would play only a limited, supporting role in nuclear decision-making. Nevertheless, considering possible application of AI in decision-making support systems requires us to revisit the viability of some conventional wisdoms as assumptions for nuclear deterrence.

Growing Questions over Rationality Assumption

Nuclear deterrence is a situation where states seek to achieve political and security goals by influencing the other side, or adversary, without using them. Some argue that deterrence could work because the high prospect for catastrophic results in the use of nuclear weapon or the failure of deterrence would induce parties to the escalation game to become very cautious in taking actions. This logic assumes that decision makers would act rationally in crisis and seek to maximize its gain or to minimize its loss. If an adversary is a rational actor, this potential aggressor will not take actions in the first place as the aggressor knows that it would face the retaliation which would result in more harm than benefit. In the meantime, ensuring non-use depends on many factors. Among them, decision makers would seek accurate grasp of the situation, accurate knowledge or understanding on adversary's action/reaction options and calculous as much as possible, but decision makers rarely enjoy such a situation.

Nowadays, as research in behavioral economics develops, there is growing argument that casts serious doubt on the assumption of rationality in human decision-making in crisis, which strategic stability under nuclear deterrence rests on. It argues that humans cannot be counted on to always maximize their prospective gains and tend to have wrong expectations and calculations on their adversary's cost-benefit calculations. Prospect Theory tells that people will take more risk in order to defend what they have already gained but tend to be more cautious and conservative in newly gaining something of equal value. And political leaders tend to be unusually optimistic and overly confident in their ability to control events. Because of over-confidence, they may fail to cut losses and take more risks either to recover the projected loss or to regain the control. In short, people may not act in a way to maximize utility, or even not be explicitly aware of costs and benefits of certain actions that they will take (Krepinevich 2019).

This theory of the nature of human psychology may undermine the reliability of strategic stability and nuclear deterrence, but it is not unique to AI. What a scenario of introducing an AI-powered decision-making system does in this regard is to acutely depict this problem. In a sense,

arguing the potential problems of AI in decision support leads us to even more fundamental question on the rationality assumption in the nuclear deterrence logics.

So, the question is whether AI would help overcome these growing concerns on the underlying assumptions of rationality in the logic of nuclear deterrence, or it would amplify such concerns. In other words, could decision makers accept AI-supported advice/date, which is counter-intuitive to decision makers? And how do they know if adversarial decision makers take actions with or without their intuition? (See section "Fog of AI War".)

Faster and more reliable, increasingly autonomous information processing systems could reduce risks associated with the management and operation of nuclear arsenals, particularly in crisis situations. Further, as AI is super rational and free from psychological biases as well as pressure that humans are always under influence, there is a possibility that AI would be able to sharply decrease, if not eradicate, the risks of human error and misperception/misconception. If it is the case, humans may thereby achieve higher levels of strategic stability in the avoidance of accidental launch or unintended escalation.

But this argument must be carefully examined by addressing the following questions: *To which objective* should "rationality" be defined in nuclear deterrence games? *What and whose objectives* should rationality be accounted for? Is it possible to establish a decision-making process with fully informed environment? (In the first place, there is no such thing as decision-making in an environment of complete information!) Additionally, would full transparency in information on nuclear arsenals contribute to the stability in an asymmetric nuclear relationship?

The first two questions suggest how difficult it is to set goals for nuclear deterrence or strategy. Perhaps even the highest national decision makers are not clearly/explicitly aware of goals, and their goals will change as situations evolve. The other two questions point to problems that may newly arise when AI is partially employed to support decision-making systems.

Fog of AI War

In a crisis situation or in a battlefield, decision makers and commanders are suffered from the so-called "fog of war," or the uncertainty of situational awareness (von Clausewitz 1832). The "fog of war" in nuclear deterrence is a problem inherent in nuclear deterrence per se, but not exclusively inherent in AI. Adoption of AI into ISR would help clear such "fog of war" caused by the lack of information on adversary's capabilities and deployment (see section "Ability to Set a Goal"). Also, as discussed above, AI, if properly applied, could contribute to confidence building and threat reduction among nuclear-armed states. However, it would also be fair to say that AI may bring another type of "fog of war," due to its potential consequence of the introduction of AI in the decision-making process.

First, the most critical "fog of war" in the game of nuclear deterrence is the logic and reasoning that shape the intentions and preference of the adversary and decide where the red line of self-restraints is drawn. Therefore, it is unclear to decision makers where to find the equilibrium to optimize the relative gain against the adversary, which must be sought through exchange of strategic communication.

The irony of nuclear deterrence is that while construction of escalation ladders and deterrence logics are largely dependent on rationality, the irrationality of cost-benefit calculations, which is derived from fear in the human mind, can also be a constraint/restraint against escalation into the use of force in a crisis. Posing unpredictability and lack of confidence in the rational calculations of adversaries dissuades them from attacking and improves the certainty of deterrence.

Second, in the pursuit of a victory in a warfighting, AI, which "feels" no obsession with the fear of losing something or defending something presumably vital, could make a rational choice solely based on the cost-benefit calculation for a pre-set objective for combat. However, it is not certain whether this will contribute to the ultimate objective of engaging in nuclear deterrence from the perspective of managing the medium- to long-term entanglement relationships among strategic rivals, and to the satisfaction of their respective peoples.

Nuclear deterrence is a psychological game, in which the threat of using nuclear weapons restricts the actions of an adversary and manages the escalation so that the confrontation between the adversary and itself does not lead to the actual use of nuclear weapons. The situation is always changing and thus it is difficult to set clear goals to be pursued. However, presumably, decision makers may be engaged in the game of nuclear deterrence even without knowing absolute truth in the game.

Third, it is sometimes not easy for decision makers to envision the "national interest" (politically it is considered absolute truth for a sovereign state) that they intend to realize through nuclear deterrence and strategy. Moreover, it is very difficult to gauge the adversary's intentions, and it is possible that even the adversary itself does not consciously understand its strategic goals.

In this regard, it seems that the affinity is missing between characters or strengths of (narrow) AI and the required skills for decision-making during a nuclear crisis and escalation game. In a situation where strategic goals are constantly changing, narrow AI does not necessarily play a role in clearing the "fog of war." Rather, the problem of blackboxing decision-making in AI as described below creates an AI-specific "fog of war."

AI as Black Box

For decision makers who are responsible for the consequences of their decisions, a critical question is how and to what extent they can trust AI. It is probably more phycological than technical question. But can decision makers confidently choose an option, following suggestions or advice provided by AI, whose calculation process is in "black box" to decision makers? There are reports on the so-called "overlearning" problems, which have caused racial discrimination and other social problems due to its solely technical nature of processing information. Probably for the algorithm that drew such a conclusion, pure data processing resulted in such a problem. However, goals sought by decision makers also inevitably involved more social and political considerations. In these cases, AI failed to incorporate such factors in drawing conclusions, and imported "social" mistakes. In such a situation, to what extent can human decision makers be assured that algorithms are not making misinterpretation, miscalculation, or misrepresentation of the reality of the situation? Research efforts have already begun. U.S. Defense Advanced Research Projects Agency (DARPA) has started research on increasing the visibility of the rational for AI's decisions under the Explainable AI program.

Decision makers must be held accountable for the consequence of their decisions even if they rely on "advice" by AI, which may be false. It is certainly possible to prescribe the responsibility of policy makers within a legal theory, but ethically and practically, commitments to the use of weapons based on the advice by AI, which lacks the traceability, may put decision makers in a rather ambiguous position in terms of both the trustworthiness of AI advice and the readiness to assume responsibility and accountability. As studies of Behavioral Economics suggest, humans have a poor intuitive grasp of probability and inconstant expectation on cost-benefit calculation, subject to the situation (e.g., "gambler's fallacy," see, for example, Tune 1964 and Oppenheimer and Monin 2009). When AI draws a conclusion which is different from an intuition that policy maker has, can policy maker follow a suggestion by AI without knowing the logic behind AI's advice? This situation may pose another type of risk/concern to the rationality assumption of nuclear deterrence.

Another black box is adversary's AI employment policy. When decision makers do not know whether or to what extent the adversary's decision depends on AI, even only with the "prospect" for the advancement of AI and its contribution to such an ability, when it is perceived by the adversary, it would have the impact on decision maker's consideration and calculation. Since AI-enhanced ISR will expedite the targeting process, thus providing the adversary with stronger time-pressing pressure for decision to counter, the lack of information on the adversary's adoption of AI into the nuclear weapon system would amplify mistrust and paranoia between adversaries.

Also, the AI-supported decision-making process, which lacks the traceability, raises the risk and vulnerability in information security, particularly against misinformation and deception. When algorithms perceive and interpret information in a wrong way or in a wrong context and thus provide biased solutions, self-learning mechanism would reproduce the biases of the data in an accelerated pace. In this regard, defending the command-and-control system from cyberattack and countering disinformation are even more critical for the AI-supported decision process.

Consequently, "black box" phenomena in AI may increase the possibility of miscalculation, and the temptation to first use nuclear weapons before destroyed. Scharre says that the real danger of an AI arms race is not that any country will fall behind its competitors in AI, but that the perception of a race will prompt everyone to rush to deploy unsafe AI systems (Scharre 2019).

AI and Changing Characters of Nuclear Deterrence

Impact on ISR

One of first applications of AI in nuclear weapons could be a Russian underwater nuclear drone. Russia is developing a nuclear-propelled underwater drone to carry a thermonuclear warhead. Given the difficulty in communicating underwater, this "Oceanic Multipurpose System Status-6" needs to be equipped with a highly autonomous operation system, presumably supported by AI (Weintz 2018). If it would actually be deployed, it would increase the survivability of Russian retaliatory nuclear capability and improve the credibility of nuclear deterrence. Then it will inevitably trigger reactions by the United States and other states exposed to increased vulnerability with this new Russian nuclear asset.

Realistically, at the current level of AI competence, an imminent question is how to assess the impact of AI in intelligence, surveillance, and reconnaissance (ISR), which subsequently affect the perception on survivability and credibility of nuclear deterrent as well as its usefulness in maintaining nuclear stability and threat reduction.

During the Cold War, the development of ballistic missile systems significantly shortened the time to deliver nuclear weapons. Sophistication of delivery systems required nuclear-armed states to develop a kind of automation and standard operating procedure to respond and retaliate adversary's nuclear attacks in a timely and effective manner. The warning time for ballistic missiles was so short that launch-on-warning postures with detection and early warning systems were also required to be established and maintained.

In order to support the operation of such systems, robust communications, control and response systems to integrate information from various sources were also constructed. Operating such complicated weapon systems effectively and credibly entails minimizing the risk of misinformation (and subsequent false alarm), misinterpretation of information, mechanical errors of early warning systems under the very strong time pressure as such misconducts might lead to a catastrophic consequence. Vulnerable, time-critical targets remains a source of serious concern and it may be even more critical in the AI-enhanced environment.

In today's warfighting domains regardless of conventional or nuclear, much information is collected and analyzed using various tools at both the strategic and theater levels. Ironically, this leaves military analysts and decision makers in a state of overabundance of information. Given its strengths in data and imaginary processing and anomaly detection, AI along with sophisticated sensing technology would make the huge difference in ISR capabilities (Kallenborn 2019).

To respond such a situation, U.S. Department of Defense launched a project called Project Maven (2017), in order to "reduce the human factors burden of [full-motion video] analysis, increase actionable intelligence, and enhance military decision-making" in the campaign to fight against ISIS. This project demonstrates the potential of AI in enabling targeted strikes with fewer resources and increased accuracy/certainty (Loss and Johnson 2019).

AI-enhanced ISR would provide with the improved ability to find, identify, track, and target their adversaries' military capabilities and critical assets. It would allow for more prompt and precise strikes against time-critical targets such as ground-based, transporter-erector missile launchers and submarine launched ballistic missiles, which are platforms to ensure the survivability of the second-strike capabilities as deterrent forces. If a country acquires exquisite counter-force capability along with a credible, AI-enhanced ISR capability, it will not only be able to limit damage in the event of a nuclear crisis escalation, but will also be able to neutralize enemy nuclear deterrence. It affects the calculation on strategic stability, which is based on the survivability of secure second-strike nuclear forces.

Whether it would contribute to enhancing the stability or undermining it, experts' views are divided.

One argument is that the AI-enhanced ISR capability could increase stability as it provides nuclear-armed states with better information and better decision-making tools in time-critical situations, reducing the risk of miscalculation and accidental escalation. Another merit of stronger ISR capabilities is the possible improvement of monitoring nuclear weapon-related developments and conducting monitoring and verification operations, which supports the compliance of arm control and disarmament arrangements (if any).

If such "AI revolution" in ISR happens equally on all nuclear-armed parties who are engaged in mutual deterrence and seeking a point of "strategic stability" (while assuming that it is no longer viable to consider only US-Russia strategic stability for the nuclear stability at the global level), monitoring and transparency on nuclear assets and activities would be increased, and the high level of verification of arms control arrangements would become possible. They would eventually improve mutual confidence among these nuclear-armed states, and contribute to threat and risk reduction among nuclear-armed states.

Another argument is that the risk may increase if such "AI revolution" in ISR happens asymmetrically, especially if the emulation of technology occurs unevenly. The risk of uneven emulation of the technology is not negligible as AI has great impact in ISR capabilities. Uneven emulation of the technology would bring a gap in ISR capability and then counter-force capability.

In this situation, AI could undermine deterrence stability and increase the risk of nuclear use for the same reasons of enhancing security. If one country would be confident in its superiority, it considers that any conceivable gains from the use of force including nuclear weapons outweigh the cost and damage caused by adversary's retaliation. AI is an enabler for gaining the superiority. On the contrary, when facing a nuclear-armed adversary with sophisticated technology and advanced ISR capabilities (if it is known), such poor performance would prove disastrous. One with weaker ISR capabilities may become concerned about the survivability of its nuclear capabilities, and may be tempted to strike before its capabilities are attacked in time of crisis. It is a classical security dilemma situation.

There is also a possibility that states, which suffer adversary's first-mover's advantage, may be tempted to offset by another means rather than catching up in the same domain. For example, when the United States demonstrated its capability for precision-strike, combined with sophisticated ISR system during the Gulf War in 1990–1991, Russia perceived the increased risk of losing a conventional war or the vulnerability of its nuclear arsenal against precision attacks. What Russia did to offset this technological disadvantage was to develop low-yield nuclear weapons and to employ a nuclear doctrine to use such weapons. If one side acquires the superiority in ISR systems powered by AI algorithms and sensing technology, and improves its ability to locate and target nuclear-weapon launchers and other strategic objects, whereas the other side's policy options are either to catch up in the technology, or lower the predictability of its behavior and make the cost-benefit calculation of nuclear attack more complicated, it may result in the destabilization of strategic relationship.

In this situation, states, which employ minimum nuclear deterrence, would be more affected by such asymmetrical

development and adoption of AI into ISR systems. They will be incentivized to expand its nuclear arsenal both in number and in variety of launching platforms in order to cope with the vulnerability. Or, they would reconsider their nuclear doctrine by raising alert status, and lowering the threshold for nuclear use by automating nuclear launch and/or by employing first use policy, which might increase the risk of escalation or triggering a nuclear war by misjudgment of the situation, and lower the possibility of avoiding accidental or inadvertent escalation.

Another factor to complicate the calculation is the accuracy and trustworthiness of AI. In theory, AI-based image recognition systems could identify second-strike capabilities. But Loss and Johnson (2019) highlight two key challenges: bad data and an inability to make up for bad data. It may not be impossible to distinguish between a regular track and a mobile missile launcher in satellite images as image data of adversary's mobile missile launchers is not sufficiently available for comparison. Further, it is possible for the adversary to take advantage of the characteristics of data processing of AI and avoid detection or input false information to deceive AI. This is a very risky practice that increases the likelihood of unwanted attacks, but at the same time, the likelihood of such misinformation may be a factor that makes it difficult to rely on AI-based ISR to make decisions. (However, it is also true that this view depends on the other party's expectation of rationality.)

While narrow AI could achieve near-perfect performance for assigned mandates in ISR and thereby enable an effective counter-force capability, inherent technological limitations and human psychological boundaries will prevent it from establishing stable deterrence relationship. AI may bring modest improvements in certain areas, but it cannot fundamentally alter the calculus that underpins deterrence by punishment.

Challenges for Stably Controlling Nuclear Risks: Arms Control and Entanglement

As seen above, AI could improve the speed and accuracy of situation awareness by utilizing neural networks, imagery sensing technology, and a huge database. Algorithms and control systems for "swarming" autonomous weapon systems, hypersonic weapons, and precision-guided weapons, as coordinated with various military assets including early warning systems, could make a huge difference in battlefield. Coordination through layers of algorithms also work to help manage complex operation. Applications of the technology in these ways to the conventional weaponry systems may change the character of war (Acton 2018). When AI assists decision makers and field commanders in choosing optimal battle plan, it would inevitably transform force structure and employment, as well as organizational and operational modality in command and control systems in order to catch up with the rapid pace of changing situations. Combined with new technologies such as precision-guided missiles and hypersonic gliders, the emerging war eco-system could heighten the vulnerability of strategic assets including nuclear assets against non-nuclear strategic weapons, and drastically shorten decision-making time to respond attacks against them.

The application of emerging technologies such as hypersonic gliders, robots, along with AI, to weapons has increased the strategic value of non-nuclear weapons. Thus, the boundaries between nuclear and conventional weapons and between strategic and non-strategic weapons have become blurred. This has increased the complexity of "cross-domain" deterrence calculations, and the vulnerability of infrastructure supporting deterrence, such as cyber and space, has made it difficult to establish the scope of an arms control regime for managing stable strategic relationships.

Historically, in order to avoid unintended and unnecessary conflicts and escalation and to maintain stable relations (maintaining arms race stability and crisis stability), adversaries have established arms control regimes. In order for an arms control system to contribute to the stable control of strategic relations between nuclear powers, it is necessary to establish a mutual understanding concerning the definition of the state of stability in a tangible manner. And the stability is often converted into the balance of forces (like "strategic stability" between two major powers). In other words, the arms control regime does not mean a pure balance of power based on an estimate of military power, but it means institutionalizing a relationship formed by mutual recognition of the "existence of equilibrium" and its joint understanding for stable management of the situation within a certain range.

Here are some key questions: Is arms control over the employment of AI in nuclear forces possible? Would AI contribute to establishing a stable arms control regime? AI is a software-based technology that makes a tangible assessment of its capabilities difficult. It suggests that an arms control regime for AI or the verification of the functioning of AI in weapon systems would be neither possible nor credible. Nuclear-armed states could therefore easily misperceive or miscalculate to what extent they should count on the impact of AI in their adversaries' capabilities and intentions. AI also help enhancing values of non-nuclear weapons used for strategic objectives rather than battlefield warfighting. It implies that designing arms control scheme by category or type of weapons may become less relevant to achieving a stability between adversaries.

In the field of nuclear strategy and deterrence, the perception of an enemy's capability matters as much as its actual capability. A worrisome scenario would be a situation where a nuclear-armed state would trigger destabilizing measures

(e.g., adopting new and untested technology or changing its nuclear doctrine) based only on the belief that its retaliatory capacity could be defeated by another state's AI capabilities (Boulanin 2019).

Agenda for Nuclear Ethics in the AI Era

Ability to Set a Goal

As seen in severe accidents that have occurred in clinical decision-support systems, aviation, and even nuclear command and control in the past, excessive reliance on automated systems (automation bias) could become a cause of error. The point here is that it is not a machine that makes mistakes, but the humans who misuse or abuse the system. An AI-enhanced decision-making system may have to be operated in a short, time-constrained fashion, which may not permit human decision makers to conduct re-evaluation and review of conclusions/advice that an AI-enhanced system provides. In this situation, the risk of automation bias would be even greater.

And even with AI support, there are so many factors to consider in decision-making that there are unconsciously many ethical and normative constraints. (Also see above the section "Challenges for Stably Controlling Nuclear Risks: Arms Control and Entanglement" for discussion on the limitation of AI's capability in "autonomous" strategic decision-making.)

Of course, these ethical and normative constraints are likely not universal, and there is no clear understanding of the extent to which they are common in different sociocultural contexts or impose constraints on decision makers. Will AI algorithms be able to identify and learn about patterns and frameworks of thought that humans do not consciously recognize? With the current level of technological competence, the limitation of AI in decision-making is clearly shown in this point.

From an ethical point of view, too, it is unlikely or unimaginable that humans will not be involved in decisions about the use of nuclear weapons. While global/universal human interests are potentially recognized as absolute good in concept, they are not prescriptive enough to serve as operational grounds for policy implementation. In the current international system, where sovereign states are major players, states are supposed to maximize their individual "national" interests. In this context, a norm of the prohibition of the use of nuclear weapons has not gained the universality, and the use of nuclear weapons is considered as a possible option for some states in an extreme circumstance of state survival.

In the meantime, the humanitarian dimension of nuclear weapons casts a doubt on the legitimacy of any use of nuclear weapons. Even among those who support the importance of nuclear deterrence for the maintenance of international peace and security, many believe nuclear weapons should never be used. It is because that once a nuclear weapon is used, it is highly likely that its consequence would go beyond the victory in war between states and reach a point that the damage to human beings and the entire earth would be unrecoverable. Managing nuclear weapons involves consideration of the tremendous social, economic, and political costs.

Nuclear deterrence is a game to be played against this kind of premises. It is a very complicated statecraft whose goal is not so straightforward as the winning a war or destroying targets. While there is a clear value standard for the use of nuclear weapons in the context of the abstract conceptual arguments of ethics and morality, when we look at the operations of policies in the modern real world, the criteria become ambiguous.

In the foreseeable future, we can hardly imagine that AI alone would set a goal and make decision of the use of nuclear weapons on behalf of humans.

Taking the Responsibility and Accountability Seriously

Automation of a decision means that a decision is made based on prescriptive standard-operating procedures. There should be no deviation as long as the automated process would not be disrupted. During the Cuban missile crisis, as we later found, there were a couple of occasions that decision makers did not follow the standard operating procedures. So deviations from preset procedures actually happened, and they could be (or at least some interpreted them as) reasons for the avoidance of escalation into nuclear exchange. This example suggests that autonomy entails adoptability over the automated decision procedure in the evolving circumstances.

The responsibility and accountability of certain behavior is closely associated with the autonomy of the system. So, is Autonomous Intelligence really autonomous?

The human mental system is a closed system, and actions based on each other's intentions are unpredictable in an ultimate sense. In other words, "unknowability" is the basis of so-called "free will" as well as the fluctuation of semantic interpretation, and thus responsibility in behavior. Although AI may appear to have a free will like a human, it is an adaptive, heteronomous system, and it is impossible to make truly autonomous decisions.

When there are multiple options and it is not clear which one the other prefers to choose, the social effect which might be brought by the government's decision of one particular option is linked to "responsibility."

Therefore, "free will" and "responsibility" are the concepts associated with closed autonomous systems, and cannot be established with open, heteronomous systems. (However,

if the behavior of the mental system is under the constraints of the upper social system and the choice is in fact capped, there is no free will and no liability.)

The behavior of the adaptive system (current "narrow" AI) is heteronomously predefined by the designer at a more abstract level. True autonomy is tied to the unknowability for the others.

Fluctuation in the interpretation of the meaning of other party's words or deeds is fundamentally due to the unknowability of the other party, which is a closed system. Since the operation of an AI is performed based on a very complex program, it would seem from the outside that predicting its output would be practically impossible (even if possible in theory). Thus, AI gives the impression that it may have "free will."

In 2012, neural network based AI, which was developed by Google, successfully identified faces of cats from ten million YouTube thumbnails without being fed information on distinguishing features that might help identify cat's faces. This experiment shows the strength of AI in detecting objects with certain characteristics in their appearance. Google's experiment appears to be the first to identify objects without hints and additional information. Five years later, an AI was trained to identify more than 5000 different species of plants and animals (Gershgorn 2017). The network continued to correctly identify these objects even when they were distorted or placed on backgrounds designed to disorientate.

However, various concepts that we deal with in politics, economy, and other human activities are different from identifying animal's face from information on social network services. It is impossible to distinguish or identify certain concepts simply by differences in their appearances alone. Concepts are relative things that differ from one language community to another, and there is no universal absolute concept that segments the world.

In playing nuclear deterrence or strategic games, which involve highly political, abstract concepts of humanitarian and ethical values beyond mere war planning rationality, decision makers (not field commanders) must take into close consideration on so many political, social, economic, and normative factors such as freedom, rights, and ethics, in a situation where the adversary's intention is unknown (with incomplete information).

As we have witnessed the 75 years of the history of the nonuse of nuclear weapons, ethical questions on the consequence of possible nuclear exchanges affected the consideration of decision makers' employing nuclear option in strategic confrontation.

Can AI incorporate highly abstract social concepts such as freedom and rights in drawing a conclusion on policy priorities or assessment on the situation? This makes us aware of the difference between human knowledge and universal, absolute knowledge. It further leads us to a question whether in particular AI will be able to provide universal knowledge. An answer at this stage of technological development may be No. Then the question further goes; Can decision makers define and describe a goal of nuclear strategic game in a way that AI could read and operate, incorporating abstract, normative concepts? Assuming it is possible, would cultural differences in background and interpretations of these concepts be overcome in order to maintain the high level of mutual predictability and thus stability.

Conclusion

Problems associated with the application of AI into nuclear deterrence command-and-control and decision systems may not be unique to AI. Rather, AI, or AI-enhanced weapon systems amplify the risks intrinsic to nuclear deterrence.

Fast and effective detection and identification of targets with AI and enhanced sensing technology would help confidence-building in one way. In another way, it poses more vulnerabilities to nuclear-armed states and increases insecurity. Space for strategic ambiguity, which in reality functions as a kind of buffer zone between deterrence and the actual use of nuclear weapons, will become narrower by AI. Fast identification and analysis of the situation may enable decision makers to consider the best option, while reaction by others may also become quicker, and allowance time for decision-making may in fact become shorter, and decision makers may have to decide and act under stronger time pressure. Therefore, prisoners' dilemma and chicken game situations in nuclear deterrence may take more acute modalities in the AI-enhanced security environment.

We will not likely see a world where humans are completely replaced by AI in nuclear decision-making in the foreseeable future. Nor is it realistic that AI would be totally dismissed from the operation of nuclear arsenal. The U.S. Department of Defense emphasizes the concept of human-machine teaming: Humans and machines work together symbiotically. Humans provide higher-order decision-making and ensure ethical and appropriate operation of autonomous systems (Kallenborn 2019).

Examining the applicability of AI to managing nuclear strategy and deterrence raise the awareness of the necessity to re-examine the understanding and appropriateness of the traditional, long-overdue question in detail, that is, whether assumption of rationality and ambiguity in the logic of nuclear deterrence is appropriate.

If AI is to give us a chance to face these hard questions on nuclear deterrence, AI may save humanity. But without addressing these concerns discussed above, AI may move forward the doomsday clock, and make us closer to nuclear winter.

References

Acton, J. M. (2018). Escalation through entanglement: How the vulnerability of command-and-control systems raises the risks of an inadvertent nuclear war. *International Security, 43*(1), 56–99. https://doi.org/10.1162/isec_a_00320. Retrieved February 25, 2020.

Boulanin, V. (2019, May). *The impact of artificial intelligence on strategic stability and nuclear risk*. Sweden: SIPRI, Stockholm International Peace Research Institute.

Geist, E., & Andrew, J. J. (2019). *How might artificial intelligence affect the risk of nuclear war?* Santa Monica, CA: RAND Corporation.

Gershgorn, D. (2017). *Five years ago, AI was struggling to identify cats. Now it's trying to tackle 5000 species*. Available via Quartz. Retrieved February 25, 2020, from https://qz.com/954530/five-years-ago-ai-was-struggling-to-identify-cats-now-its-trying-to-tackle-5000-species/

Heath, N. (2018). *What is machine learning? Everything you need to know*. Available via ZDNet. Retrieved February 25, 2020, from https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/

Jervis, R. (1989). *The Meaning of the Nuclear Revolution: Statecraft and the Prospect of Armageddon*. Ithaca, NY: Cornell UP.

Kallenborn, Z. (2019). *AI risks to nuclear deterrence are real*. Available via War on the Rocks. Retrieved February 2020, from https://warontherocks.com/2019/10/ai-risks-to-nuclear-deterrence-are-real/

Kissinger, H. A. (2018). *How the enlightenment ends. Available via The Atlantic*. Retrieved February 25, 2020, from https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/

Krepinevich, A. F. (2019). *The eroding balance of terror: The decline of deterrence*. Available via Foreign Affairs. Retrieved February 25, 2020, from https://www.foreignaffairs.com/articles/2018-12-11/eroding-balance-terror

Loss, R., & Johnson, J. (2019). *Will artificial intelligence imperil nuclear deterrence?* Available via War on the Rocks. Retrieved February 25, 2020, from https://warontherocks.com/2019/09/will-artificial-intelligence-imperil-nuclear-deterrence/

Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making, 4*(5), 326–334.

Podvig, P. (2012). *The myth of strategic stability*. Available via Bulletin of the Atomic Scientists. Retrieved February 25, 2020, from https://thebulletin.org/2012/10/the-myth-of-strategic-stability/

Scharre, P. (2019). *Killer apps: The real dangers of an AI arms race*. Available via Foreign Affairs. Retrieved February 25, 2020, from https://www.foreignaffairs.com/articles/2019-04-16/killer-apps

Schelling, T. (1966). *Arms and influence*. New Haven, CT: Yale UP.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Science, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756.

Tune, G. S. (1964). Response preferences: A review of some relevant literature. *Psychological Bulletin, 61*, 286–302.

von Clausewitz, C. (1993) *On War*. London: Everyman. Originally published as *Vom Kriege* (in German) in 1832.

Weintz, S. (2018). *The real reason you should fear Russia's status-6 torpedo*. Available via The National Interest. Retrieved February 25, 2020, from https://nationalinterest.org/blog/buzz/real-reason-you-should-fear-russias-status-6-torpedo-28207