The guided-missile destroyer USS *Porter* conducts strike operations against a target in 2017. (U.S. Navy/Ford Williams)

# The Machine Beneath

## Implications of Artificial Intelligence in Strategic Decisionmaking

By Matthew Price, Stephen Walker, and Will Wiley

On the morning of May 17, 2024 U.S. and Chinese leaders authorized a limited nuclear exchange in the western Pacific. No one, including those who made the decision, is completely sure what caused the flash war. However, historians are confident that neither side deployed fully autonomous weapons or intentionally violated the law of armed conflict. Each side acted in anticipatory self-defense. Irrespective of the intent, in less than two hours, the technologies in use prompted a conflict that killed millions.

### The Two Hour Road to War

As the 2020s dawned, tension in the South China Sea (SCS) continued to escalate. China, the United States, Vietnam, the Philippines, and Taiwan expanded their deployment of forces in the region and made increasingly uncompromising territorial claims. At the same time, both the United States and China fielded artificial intelligence (AI) systems designed for strategic, multi-domain awareness. AI allowed the fusing and incorporation of billions of diverse data points across a spectrum of intelligence sensors, online meta-material, social media, economic indicators, and human reporting. Initial versions of the U.S. systems received extensive verification and validation of performance to ensure system reliability and safety assurance. These systems proved highly reliable in identification of the various predictors of adversary actions around the globe, as did the Chinese systems, despite receiving less rigorous testing and evaluation. AI development proceeded rapidly and by late 2022 dual-use AI systems became commercially available that could not only synthesize fused data, but make recommendations for complex problem sets. As an early adopter of "strategic AI" for military decisionmaking, Israel acted upon a strategic AI recommendation to outmaneuver and defeat an impending Syrian attack in February 2023.

With tension reaching new heights that summer, leaders in China and the United States clamored for immediate fielding of the latest strategic AI to gain faster analysis and advantage over potential adversaries. The newest system was not only trained on historical engagements, wargames, and

At the time of writing, Lt Col Matthew Price, USAF, was a political-military affairs officer in the Joint Staff Directorate for Strategy, Plans, and Policy (J5). LTC Stephen Walker, USA, was an exercise and training chief for the Office of Defense Cooperation in Greece. CDR Will Wiley, USN, was the Commanding Officer of the USS *John Warner*. National Defense University's Joint and Combined Warfighting School selected their article for the MacArthur Foundation Award in March.

intelligence data, but drew primary insight from competing against versions of itself, generating novel and superior strategic recommendations. Buoyed by a chorus of voices expounding the immediacy of the threat, and the exemplary performance of prior versions, in January 2024 the United States fielded the Enhanced Autonomous Recommendation System (EARS), with test and evaluation still incomplete. Senior leaders mitigated the risk by ensuring that no lethal weapons system could be autonomously authorized by EARS and each system recommendation was provided to a human decisionmaker. China fielded a similar system, the Advanced Military Advisory System (AMAS) and likewise ensured it could not authorize autonomous lethal action. However, in order to gain an asymmetric advantage, China directly linked non-lethal systems to AMAS's recommendations, allowing preparatory activity in integrated logistics, cyber, space assurance, and energy management. Acting on an AMAS decision, these systems would autonomously increase levels of preparedness.

On May 17, 2024, at 0435 local time in the SCS, a Vietnamese registered fishing vessel collided with a Chinese flagged vessel en route to a manmade island. The collision caused the Chinese vessel to sink along with its classified sensor and hypersonic weapons package. Ten minutes later, a Carrier Strike Group—the largest operational unit of the U.S. Navy—entered the region. Vietnamese forces had been preparing during the last few days for ground exercise maneuvers north of Hanoi, within 100km of the Chinese border. A few hours earlier, markets in the United States reacted to the unexpected cancellation of a Malaysian hedge fund's purchase order from the China National Offshore Oil Corporation, which extracts from the SCS. At 0439, social media hits for an anti-Chinese Government demonstration, initiated early that week by a half-American Chinese dissident, exceeded a predefined government threshold of ten million hits. At 0447, EARS notified U.S. Indo–Pacific Command (USINDOPACOM) that Chinese cyber intrusions had spiked on Pacific networks, automated PRC logistics systems were activating, Chinese national space assets had initiated defensive orbital maneuvers, and power generation had spiked at Chinese railgun sites. In addition, EARS assessed an impending major internal security operation in Beijing and noted the embarkation of numerous Chinese naval vessels to the SCS.

By 0500, USINDOPACOM headquarters, the Pentagon, and the White House were deliberating a response. At 0505, AMAS notified Chinese military leaders of increased communication loading between Pacific sensor sites, USINDOPACOM, and key command and communication nodes in Washington. EARS assessed the Chinese might be considering a lightning strike against allied-held islands in the SCS. EARS recommended an elevated posture to deter this action and at 0515, U.S. leaders authorized an immediate increase in threat level and issued orders for a strategic show of force. Incorporating indications of this U.S. action, AMAS assessed the United States was preparing a strike to eject China from the SCS and contain long-term Chinese military expansion. AMAS recommended achieving escalation dominance through a limited, preemptive strike against major American, Vietnamese, and Philippine bases using high-velocity weapons including cyber, ballistic, hypersonic, and counter-space assets. Believing the U.S. attack would commence at any time, at 0520, Chinese leaders authorized the strike, for which AMAS had already calculated and prepared the plan of action. At 0520:07, EARS alerted U.S. leaders that American forces were under massive cyberattack and Chinese hypersonic missiles would impact the Philippines in 73 seconds. EARS recommended a limited nuclear strike to end the Chinese attack, assessing that the Chinese would only respond in kind, but avoid an attack on the U.S. homeland and subsequent strategic annihilation. U.S. missile defenses could successfully intercept a significant portion of the Chinese theater nuclear response. EARS was correct. By 0624, after a limited nuclear exchange, millions were dead, tens of millions wounded, and both sides had ceased hostilities.

During peace talks, the Chinese claimed their systems were only responding to Vietnamese maneuvers, U.S. war preparations, and the suspicious sinking of a Chinese ship. The United States claimed it was only reacting to aggressive Chinese actions. Secretly both sides attempted to reconstruct in detail the analysis and decisions made by EARS and AMAS. However, the creators of the underlying AI systems explained that it was not possible to keep retroactive interrogation protocols up to date under the time constraints imposed by the military and business customers. These protocols are a data reconstruction program designed to explain the decision rationale of the AI. Even with those protocols, knowing the reasoning behind every subset decision was likely not possible given the system design. Though EARS and AMAS were decommissioned, the advantages of strategic AI proved too valuable to fully reject, as Russia and

In March, an unarmed Trident II D5 missile launches from the USS *Nebraska* in the Pacific Ocean. The successful test launch certified the readiness of the ballistic submarine missile crew and the operational performance of the submarine's strategic weapons system. (U.S. Navy/ Ronald Gutridge)

Iran continued to advance their systems, and today almost all nations field these systems.

## Did AI Cause the Flash War of 2024?

This fictional narrative demonstrates how two nations might enter into war without their respective leaders fully understanding why their countries were engaged in hostilities. The speed of future military action will incentivize increased reliance on complex autonomous analysis, recommendation, and decisionmaking. Autonomous systems of today are limited by an inability to synthesize large data sets in a complex operating environment, but future systems are likely to possess such capability. How will these technologies be managed and implemented in the future? Will they be treated with the same caution and control as strategic nuclear weapons or seen as just another innovative tool for understanding the military environment?

To date, the debate about fielding autonomous military systems has focused on ensuring a human is in the decision loop for lethal engagements by battlefield weapons.[1] Mostly absent is a rigorous debate over the implications of relying on autonomous systems for strategic analysis and recommendation. AI was only used to recommend action. Leaders in the loop chose the actions that led to a nuclear exchange. The decisionmaker trusted the recommendation of the strategic AI similar to the trust placed in map applications used to find the fastest route to a destination. Individuals do not fully understand how the route is generated, but they know it is normally correct, so they follow it. Such reliance on autonomous systems for strategic recommendation could both marginalize the value of human review and invalidate the historical assumptions about adversary intent that underpin deterrence theory, increasing the risk of escalation. U.S. policy and international law must take these risks into account in developing new norms for militarized autonomy, broadening their scope beyond battlefield autonomy to AI systems created to support strategic decisionmaking, or what these authors term strategic AI.

## When We say AI what do We Mean?

Understanding the scope of AI under consideration is critical to addressing the potential impacts of reliance on increasingly sophisticated AI for military decisionmaking. In lay terms, AI breakthroughs can conceptually be categorized into three categories:

- Artificial Narrow Intelligence (ANI) already exists. Examples include image recognition and game supremacy in chess or Jeopardy;
- Artificial General Intelligence (AGI) is a human level intelligence that operates across a similar variety of cognitive, creative, and possibly emotional tasks. Often the focus of intellectuals and futurists, this level of AI does not exist;
- Artificial Super Intelligence (ASI) is another theoretical category of artificial intelligence beyond any human level. Notably, achieving ASI is not required for AI to be superior to human intelligence in particular tasks.[2]

This article is concerned with the application of likely improvements in ANI, which will allow for automated synthesis and analysis of large, diverse data sets, and provide utility in military decisionmaking, as envisioned in the Flash War of 2024. Numerous avenues within AI are under development, with recent, impressive gains in machine learning and artificial neural networks.[3] This article does not attempt to predict the future of specific AI mechanisms, but rather grapples with general implications of applying advances in the broad field of ANI to military strategy.

## Ambitions for Military Use of AI

The U.S. National Defense Strategy (NDS) declares that future defense spending will focus on AI:

> *The Department will invest broadly in military application of autonomy, artificial intelligence, and machine learning, including rapid application of commercial breakthroughs, to gain competitive military advantages.*[4]

The pursuit of AI development and research outlined by the NDS has been matched with dedicated funding in the 2019 Pentagon budget submission.[5] These are overt messages to our near-peer competitors and adversaries, but there is no clear discussion or policy on how this money will be spent in the outlying years.

The United States' main competitors, namely China and Russia, are also investing heavily in AI. In July 2017, China released its "New Generation Artificial Intelligence Development Plan," which outlined efforts to "lead the world" in artificial intelligence by 2030.[6] The leadership in Beijing is serious about leveraging breakthroughs in the private sector and utilizing them in a military capacity. Likewise, Russian President Vladimir Putin said recently, "Artificial intelligence is the future not only of Russia, but of all of mankind."[7] From these statements and actions, it is clear that the United States' main competitors are investing heavily in this technology but, as for the United States, how they will use this technology remains unclear.

## U.S. Military Guidance and Attempts to Establish International Norms

Department of Defense (DOD) Directive 3000.09 "Autonomy in Weapon Systems" signed in 2012 by then Secretary of Defense Ashton Carter serves as the governing document for AI technology in DOD.[8] The directive aims to "minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements" and to ensure there are "appropriate levels of human judgement over the use of force." The U.S. military must remain compliant with the Law of Armed Conflict (LOAC) principles of military necessity, distinction, and proportionality. The primary concern with lethal autonomous weapons systems (LAWS) is whether they can determine if a target has the military significance (military necessity), that the target is a combatant (distinction), and if the military action is overkill to accomplish the task (proportionality).[9] To date, the U.S. military is worried about whether or not the LOAC and the rules of engagement (ROE) for the operation can be adequately applied by an autonomous system without a human in the "loop" to make these decisions. It is unclear if China or Russia have a similar policy.

Many civilian and government parties have called for international agreement on LAWS. Beginning with informal discussions in 2014, the UN Convention on Certain Conventional Weapons (UNCCW) has addressed LAWS.[10] In 2017 these discussions were elevated to the UN Group of Governmental Experts process wherein the majority of GGE members agreed on the need for a legally binding instrument of regulation and the need for meaningful human control over LAWS.[11] The 2017 and 2018 GGE meetings set the framework for further discussions, but few concrete steps toward formal agreement have been made. In 2017 Russia announced it would not adhere to any "international ban, moratorium, or regulation on such weapons."[12] In concert with the 2018 talks, 26 countries pledged support for a ban on fully autonomous weapons, including surprise support by China.[13] However, key nations including the United States, opposed negotiating a politically or legally binding document on LAWS. The U.S. delegation at the GGE argued that autonomous weapons development should be consistent with existing International Humanitarian Law and a ban would constitute an unrealistic and premature judgement on the future of autonomous technologies.[14] As evidence of the complexity of the problem, the international community has yet to agree on a detailed definition of LAWS. Like the DOD guidance, international discussions have mostly focused on tactical systems.

## How Autonomous is the U.S. Military Arsenal?

The United States already fields autonomous technology on the battlefield. For example, the U.S. Navy's submarine force uses the Mk 48 advanced capability torpedo that has autonomous features. A human makes the decision to launch the weapon at a target and takes great pains to enter search parameters into the weapon that will avoid engagement of a non-combatant vessel. However, if the weapon detects another target inside of its search parameters it will home in and destroy that object without checking with the human. It makes the "decision" to lethally strike an unintended target despite the best efforts of the operator who fired the torpedo.

Similarly, the U.S. Navy's Close-In Weapon System (CIWS) has a full autonomous mode. This system is a rapid firing gun mounted on many of the Navy's surface warships that can detect incoming aircraft or missiles and shoot them down before they attack or impact the ship. A human decision-maker determines when the system is placed in automatic, but once it is in that mode the system engages anything it detects as a target whether it is friendly or enemy. Like the torpedo, the U.S. Navy trusts the system operators to employ the CIWS in automatic mode only when the targets detected are likely to be enemies.

Finally, the Missile Defense Agency's Counter Rocket Artillery and Mortar (C–RAM), is used to protect U.S. military bases and sensitive areas. C–RAM aggregates fire finder and aviation radars, and acoustic sensors to verify incoming indirect fire in fractions of a second. The system verifies the safety of friendly aircraft by comparing their locations with the potential effects of engaging the incoming rounds. Optional responses include localized alarms to personnel and direct fire interceptors to shoot down the incoming round mid-air. C–RAM has an autonomous mode where it fires without additional permission at an incoming missile it detects.

These are just a few of the examples in the current U.S. arsenal that feature autonomous capabilities. These systems, and their inherent risks, are accepted by senior leadership as not violating the LOAC treaties because they have been fielded for numerous years (torpedo and CIWS) or are defensive in nature (CIWS and C–RAM). Policymakers appear unwilling to change the standing policy, believing the system is "working" since no treaties are violated and a human is in control of the technology.

Though the current debate about LAWS centers on AI-empowered advances in the autonomy of such battlefield weapons, the use of AI for strategic decisionmaking may pose a greater threat.

## AI Evolution and Adaptation, and the Changing Character of War

In 2017 the UN Institute for Disarmament Research used the term "flash war" to describe shortened time cycles between decision and action, incentivized by increasing autonomy.[15] The Flash War of 2024 is a creative extrapolation of emerging AI capabilities, applied to the strategic environment. If this prospect seems unrealistic, consider that in March of 2016, Google's AlphaGo AI made headlines with an unprecedented defeat of world champion Lee Sedol at the extremely complex board game Go.[16] It accomplished this feat after months of exhaustive observation, practice against human experts, and play against versions of itself. The next year, Google programmers unveiled AlphaGo Zero. In only three days it exceeded the abilities of the version that defeated Sedol, and did so without any historical knowledge, observation, or human intervention.[17]

Hedge funds, such as the Man Group, have been using similar technology for years to analyze large financial data sets, along with sources such as press releases, corporate reports, and other information.[18]

According to an interview by Bloomberg Market, the CEO of the Man Group, Luke Ellis, admittedly quarantined the system when it was first introduced because its engineers could not "explain why the system was executing the trades it was making."[19] Notably, the Man Group's incorporation of the software is reported to have taken off after test runs demonstrated consistent profit making, not as a result of a method for explaining all the trading decisions.[20] The Man Group now has AI incorporated into management of more than $12 billion, and many other investment groups have followed suit, despite the fact AI analytics are known to occasionally drive bad investments based on spurious correlations.[21]

The exact processes by which advanced AI like AlphaGo Zero come to decisions is very difficult to determine, sometimes impossible, and their complexity is increasing rapidly.[22] Though Google's engineers, and other AI researchers, are now working techniques to interpret the broad outlines of how their machines arrive at conclusions, experts such as Uber's Jason Yosinski admit that as AI ". . . get more complicated, it is going to be fundamentally difficult to understand why they make decisions."[23]

The ability to automate data analysis across multiple domains will allow a state to gain an informational and temporal advantage, through enhanced understanding of the social, political, economic, and military factors affecting a strategic environment. AI is already helping intelligence organizations sift through the data noise to find the proverbial signal.[24] The Army's Communications-Electronic Research, Development, and Engineering Center (CERDEC) has produced an Automated Planning Framework prototype to analyze the decisionmaking process, with the acknowledged goal of AI that "helps us understand, plan and fight in multi-domain battle." According to Lisa Heidelberg, CERDEC's Mission Command Capabilities Division chief, autonomy will shorten the decision process and "facilitate recommendations and predictions."[25] Separately, AI researchers have made gains in the synthetic modeling of human belief architectures to more accurately predict adversary behavior within a military hierarchy operating in simulated battlefield conditions.[26] Fusing these breakthroughs to enable strategic prediction and recommendation is the next logical step and this increasingly appears feasible in the near-term.

*The ability to automate data analysis across multiple domains will allow a state to gain an informational and temporal advantage, through enhanced understanding of the social, political, economic, and military factors affecting a strategic environment.*

Even with the advent of such systems, critics could argue that this is merely an evolution in the age-old attempts to reduce the fog and friction of war. Commanders have always tried to better understand the enemy and environment, and strategic AI will just be a new tool in this effort, so long as the software is not allowed to direct lethal action. Human leaders would not cede critical recommendation making to software in high-consequence scenarios. This argument misses two key points. First, people are generally poor judges of risk under complexity, tending to both underestimate the probabilities of failure and incorrectly judge risks posed by low probability, high-consequence events.[27] Second, time pressures can result in severe risk taking and decreased ability to discriminate probabilities.[28] Even a cursory review of existing applications of AI undermines the assumption that humans will retain strong control over important decisions, particularly as decision time is curtailed by technological advances. As mentioned earlier, AI systems are currently used extensively in equity trading and other investments decisions. Forty four percent of surveyed Americans would ride in a driverless car given the chance despite the limited real-world deployment of this technology.[29] Billions of users around the world rely on social media information curated by algorithms.[30]

Each of these examples relates to high-consequence events, where the common outcome is benign or positive. Algorithmic trading enabled the stock market flash crash of 2010, empowering a London trader's "spoofing" of the market, followed by several smaller, "innocent" flash crashes since, including a 700 point Dow Jones drop within 20 minutes in February 2018.[31] Autonomous vehicles have the potential to dramatically reduce driving fatalities due to human error, but ill-considered over-reliance on early versions have resulted in occupant deaths.[32] Russian influence operators utilized the algorithmic curation of social media during the 2016 election cycle to target susceptible groups for influence and spread divisive disinformation.[33] The automated trading example is again instructive because it bridges into the concept of an arms race in a time constrained environment. Algorithms that analyze and act on emergent data more rapidly than their competitors can make more money in trading. This is a classic, albeit non-lethal, arms race and prisoners' dilemma in a high-risk, high-speed environment. Errors in these systems have the potential to cause devastating consequences for the company acting alone or to the global economy when acting in concert, but for most individual companies they provide an unquestionable advantage, most of the time. Who is going to give them up first?

What makes this informational arms race different from historical examples is that the root cause of miscalculation might prove unknowable, as with some market flash crashes. The counterargument is that the seeds of historical miscalculation, particularly in military affairs, have often been unknown, lost in the fog and friction of war. However, the basis of such decisions was always knowable. If you could go back and ask the responsible person, they could tell you how they came to their decisions. With the evolving nature of modern AI, this is not necessarily the case.[34] AI may not only change the character of war, but even the nature of warfare as a contest of human will.[35] Even if interrogation protocols could be mandated for all strategic AI, possessing the capability to parse out the key decisional reasoning, it will likely be impossible to explain the complex rationale to a human decisionmaker fast enough to be useful in the future combat environment. In this way, strategic AI systems may reduce the friction of war, while increasing the fog. If fast, strategic AI also achieve strong average accuracy, this may result in leaders accepting high levels of absolute risk under time pressure, as investment firms do now, but with potentially catastrophic consequences.

From the invention of cavalry, to mechanized armor, to supersonic aircraft, improved speed of action has preoccupied military leaders throughout history. Today, military innovators are pushing new boundaries with technologies such as maneuverable hypersonic vehicles and cyber effects that breach the digital/physical divide. Speed has always constrained commanders' decision time and that trend is accelerating. This will generate huge pressure to rapidly understand and act upon information. AI holds the promise to help with this problem and creates a temptation to value the risk posed by the enemy much higher than the risk posed by algorithmic error and opacity. This could lead to the operational deployment of prototype AI systems that have not undergone rigorous evaluation and testing. Imagine Congressional hearings about a combat defeat in which the senior military leaders explain they did not use available strategic AI programs as capable as the adversary's because the programs were still undergoing rigorous testing for reliability and validity.

## Implications for Deterrence and Conflict Escalation

Deterrence is used to offset strategic disadvantage or the possibility of unacceptable consequences from military action. Simply put, it is designed to prevent war. Deterrence works through the existence of a credible threat of unacceptable counteraction and/or belief that the cost of action outweighs the perceived benefits.[36] In other words it is a state of mind that exists in an adversary.[37] For this reason, U.S. experts during the Cold War spent time analyzing the personality and psychology of Soviet leaders, and the Soviets sought similar understanding.[38] The knowledge obtained made each side believe in their ability to understand the other's rationale. Despite several close calls, deterrence appears to have worked, as the Cold War did not turn into another world war. The fundamental reliance of deterrence theory on the mind of the adversary causes today's theorists to question whether it can be applied to North Korea. The implications for deterrence policy are the reason the U.S. Intelligence Community has made a significant effort to determine whether North Korean Leader Kim Jong Un is a rational actor, who will act in accordance with traditional motivations.[39]

Miscalculation, failures of deterrence, and unintended escalation are branches of the same tree, and are all failures to understand an adversary. Hannibal thought the Romans would sue for peace after suffering a few decisive losses.[40] The United States thought former Iraq President Saddam Hussein would be deterred in his aspirations for Kuwait and he in turn thought the United States would not make an escalatory counteraction.[41] How would using strategic AI be different?

*Speed has always constrained commanders' decision time and that trend is accelerating. This will generate huge pressure to rapidly understand and act upon information. AI holds the promise to help with this problem and creates a temptation to value the risk posed by the enemy much higher than the risk posed by algorithmic error and opacity. This could lead to the operational deployment of prototype AI systems that have not undergone rigorous evaluation and testing.*

As previously discussed, trends in AI development are creating opaque systems with elements of their underlying decisions that cannot be fully parsed. Programmers set parameters and build in their desires and biases, as with prior computer algorithms, but validating the translation of the programmers' goals to system outputs is more difficult. If this is difficult for the system designers, it may be all but impossible for the adversarial party to understand. In a new world of autonomous analysis, how well designers know the "minds" they have created will be questionable and understanding the "minds" the adversary has created may prove impossible. As envisioned in the opening vignette, both the Chinese and the American AI systems produced outputs that were similar to their design goals, and arguably proportional, but the true intent of the adversary was mutually unintelligible. As the future speed of warfare decreases command decision time, this added layer of complexity in deterrence and escalation theory could increase uncertainty in strategic conflict.

## Recommendations and Conclusion

Despite the risks posed by the adaptation of AI to military affairs, the United States must seek to be at the forefront of this technology. It is unthinkable that America will cede this new territory to our competitors, such as China and Russia, who are aggressively pursuing it. Even if the United States decided to opt out of this arms race, it would have little effect, as the technologies described in this paper are inherently dual use, and the private sector around the globe will pursue them with abandon. So how can the United States, its partners, and its adversaries avoid consequences like those described in the opening vignette?

Ethicists, weapon engineers, and military leaders are already hard at work on the challenges associated with designing and deploying battlefield LAWS. With this article, the authors hope to begin a new conversation, highlighting and differentiating the risks posed by employing strategic AI in military decisionmaking, particularly as the pace of warfare accelerates. The following recommendations are only intended to start this new discussion, one that is related to the challenge of LAWS from underlying advances in AI, but distinct in its implications and approach. This is a wicked problem that escalation and deterrence theorists must address in haste, but there are existing mechanisms that could be employed.

DOD Directive 3000.09 should be expanded beyond a focus on individual weapon systems to address strategic concerns and to remove the current exemption for autonomous cyberspace systems designed to create physical effects. Additionally the directive should be updated to include additional guidance for systems likely to generate compounding effects or interactions with other autonomous systems.

Another avenue is international arms control, including some of the proposals that have been considered under the UNCCW. Arms control does not have to mean ceding the technological initiative. The United States and the former Soviet Union continued to improve their nuclear weapons knowledge and technology even while engaging in talks and nuclear stockpile reductions. While strategic AI is inherently different than nuclear warheads, key principles could be developed. It is in the U.S. interest to demonstrate leadership in the UNCCW discussions on autonomous weapons, in order to shape common principles and global norms. These might include proscribing not just the autonomous release of certain weapons, but even any recommendation for strategic attack by such systems. Weapons inspection regimes and a type of "Open Skies" treaty for militarized AI could be agreed to by the international community. Such agreement might seem unrealistic today, but probably no more so than Strategic Arms Limitations Talks or the Treaty on Open Skies would have been to U.S. and Soviet leaders in that late 1940s.[42]

It might also be possible for the international community to agree on some level of required interrogation capability for strategic AI. Increasing the transparency of the strategic AI decisionmaking process could decrease the chance of miscalculation. Communication protocols might also be developed that automate delays in escalatory iterations. The ultimate consequence of the opening vignette would likely be avoided if each system had recommended adversary communication during the various steps of its analysis. Though it is unlikely to be as simple as a flashing banner that says "pick up the red line," programming de-escalatory parameters may be possible.

Whatever the mechanism, deterrence, escalation, and arms control professionals must be an integral part of the development of AI systems used in military decisionmaking. Failure to address these threats before such systems are broadly deployed could have catastrophic consequences. Alternatively, the appropriate use of autonomous analysis and recommendations could actually reduce the chance of conflict, if used to decrease the fog of war without increasing strategic volatility. The United States needs to take a clear position on strategic AI. Simply stating that the U.S. Government is focusing on AI and investing in this technology is insufficient to address the coming disruption to global security. PRISM

## *Notes*

[1] Department of Defense, "DOD Directive Number 3000.09: Autonomy in Weapon Systems," November 21, 2012, Incorporating Change 1, May 8, 2017, available at <http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf (accessed Mar 6, 2018>.

[2] Cecilia Tilli, "Striking the Balance on Artificial Intelligence: We need to be cautious about AI. But that doesn't mean you should fear it." *Slate*, March 2, 2015, available at <http://www.slate.com/articles/technology/future_tense/2015/03/elon_musk_stephen_hawking_artificial_intelligence_the_state_of_a_i_research.html>.

[3] Matt Reynolds, "DeepMind's neural network teaches AI to reason about the world," *New Scientist*, June 12, 2017, available at <https://www.newscientist.com/article/2134244-deepminds-neural-network-teaches-ai-to-reason-about-the-world/>. See also: Will Knight, "Forget AlphaGo—DeepMind Has a More Interesting Step Toward General AI: Researchers are testing algorithms that display human-like ingenuity in learning," *MIT Technology Review*, June 14, 2017, available at <https://www.technologyreview.com/s/608108/forget-alphago-deepminds-has-a-more-interesting-step-towards-general-ai/>.

[4] "Summary of the 2018 National Defense Strategy of the United States of America: Sharpening the American Military's Competitive Edge," Defense.gov, available at <https://www.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf, 7.

[5] Brandon Knapp, "Here's Where the Pentagon Wants to Invest in Artificial Intelligence in 2019," C4ISRNET, February 16, 2018, available at <https://www.c4isrnet.com/intel-geoint/2018/02/16/heres-where-the-pentagon-wants-to-invest-in-artificial-intelligence-in-2019>.

[6] Elsa B. Kania, "Artificial Intelligence and Chinese Power," Foreign Affairs, December 5, 2017, available at <https://www.foreignaffairs.com/articles/china/2017-12-05/artificial-intelligence-and-chinese-power>.

[7] Zachary Cohen, "US Risks Losing Artificial Arms Race to China and Russia," CNN.com, November 29, 2017, available at <https://www.cnn.com/2017/11/29/politics/us-military-artificial-intelligence-russia-china/index.html>.

[8] Department of Defense, Directive 3000.09, "Autonomy in Weapon Systems," November 21, 2012, Crptome.org, available at <https://cryptome.org/dodi/dodd-3000-09.pdf>.

[9] Rod Powers, "Law of Armed Conflict (LOAC)," The Balance.Com, October 27, 2016, available at <https://www.thebalance.com/law-of-armed-conflict-loac-3332966>.

[10] United Nations, "Background on Lethal Autonomous Weapons Systems in the CCW," United Nations Office at Geneva, accessed March 6, 2018, available at <https://www.unog.ch/80256EE600585943/(httpPages)/8FA-3C2562A60FF81C1257CE600393DF6?OpenDocument>.

[11] Agence France-Presse at the United Nations, "'Robots are not taking over,' says head of UN body on autonomous weapons," *The Guardian*, November 17, 2017, available at <https://www.theguardian.com/science/2017/nov/17/killer-robots-un-convention-on-conventional-weapons>.

[12] Patrick Tucker, "Russia to the United Nations: Don't Try to Stop Us from Banning Killer Robots," *Defense One*, November 21, 2017, available at <http://www.defenseone.com/technology/2017/11/russia-united-nations-dont-try-stop-us-building-killer-robots/142734/>.

[13] Tucker Davey, "Lethal Autonomous Weapons: An Update from the United Nations," *Future of Life Institute*, April 30, 2018, available at <https://futureoflife.org/2018/04/30/lethal-autonomous-weapons-an-update-from-the-united-nations/?cn-reloaded=1&cn-reloaded=1>.

[14] Mission of the United States, Geneva Switzerland, "CCW: U.S. Opening Statement at the Group of Governmental Experts Meeting on Lethal Autonomous Weapon Systems: Opening Statement as Delivered by Ian McKay," April 9, 2018, available at <http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/300009p.pdf (accessed Mar 6, 2018>.

[15] UNIDIR Resources, "No. 6 The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics, and Definitional Approaches—a primer," *United Nations Institute for Disarmament Research*, available at <http://www.unidir.org/files/publications/pdfs/the-weaponization-of-increasingly-autonomous-technologies-concerns-characteristics-and-definitional-approaches-en-689.pdf>.

[16] Christopher Moyer, "How Google's AlphaGo Beat a Go World Champion," *The Atlantic*, March 28, 2016,

available at <https://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/>.

[17] "AlphaGo Zero: Learning from scratch," DeepMind, accessed March 8, 2018, available at <https://deepmind.com/blog/alphago-zero-learning-scratch/>.

[18] Adam Satariano and Nishant Kumar, "The Massive Hedge Fund Betting on AI," *Bloomberg Markets*, September 27, 2017, available at <https://www.bloomberg.com/news/features/2017-09-27/the-massive-hedge-fund-betting-on-ai>.

[19] Ibid.

[20] Ibid.

[21] Robin Wigglesworth, "Spurious correlations are kryptonite of Wall St's AI rush," *Financial Times*, March 14, 2018, available at <https://www.ft.com/content/f14db820-26cd-11e8-b27e-cc62a39d57a0>.

[22] Knight Will, "The Dark Secret at the Heart of AI: No one really knows how the most advanced algorithms do what they do. That could be a problem." *MIT Technology Review*, April 11, 2017, available at <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/>.

[23] Cade Metz, "Google Researchers Are Learning How Machines Learn," *The New York Times*, March 6, 2018, available at <https://www.nytimes.com/2018/03/06/technology/google-artificial-intelligence.html>.

[24] Daniel Terdiman, "How AI Helps The Intelligence Community Find Needles in The Haystack," *Fast Company*, October 24, 2017, available at <https://www.fastcompany.com/40484861/how-ai-helps-the-intelligence-community-find-needles-in-the-haystack>.

[25] Sandra Jontz, "Rethinking Cognition: AI and machine learning are shaping the U.S. military's decision-making process," Signal, September, 2017, available at <https://www.afcea.org/content/rethinking-cognition>.

[26] Sanjay Bisht, H.S. Bharati, S.B. Taneja, and Punam Bedi, *Defense Science Journal* 68, no. 1 (January 2018): 46–53, DOI: 10.14429/dsj68.11375.

[27] Amos Taversky and Daniel Kahneman, "Judgement under Uncertainty: Heuristics and Biases," *Science* 185, no. 4157 (September 27, 1974): 1124–31, DOI 10.1126/science.185.4157.1124. See also Nassim Nicholas Taleb and George A. Martin, "The Risks of Severe, Infrequent Events," *The Banker*, September 2007, 188–89, available at <http://www.fooledbyrandomness.com/thebanker.pdf>.

[28] Diana Young, Adam Goodie, Daniel Hall, and Eric Wu, "Decision making under time pressure, modeled in a prospect theory framework," *Organizational Behavior and Human Decision Processes* 118, no. 2 (July, 2012): 179–88, DOI: 10.1016/j.obhdp2012.03.005, available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375992/>.

[29] Aaron Smith and Monica Anderson, "Automation in Everyday Life: 3. American's attitudes toward driverless vehicles," *Pew Research Center*, Internet and Technology Report, October 4, 2017, available at <http://www.pewinternet.org/2017/10/04/americans-attitudes-toward-driverless-vehicles/>.

[30] Will Oremus, "Who Controls Your Facebook Feed: A small team of engineers in Menlo Park. A panel of anonymous power users around the world. And, increasingly, you." *Slate*, January 3, 2016, available at <https://www.google.com/search?source=hp&ei=EdmhWsQ0z8KPA4Gxit-gN&q=slate+who+controls+your+facebook+feed&o-q=slate+who+controls+your&gs_l=psy-ab.3.0.0i2 2i30k1.1894.4522.0.5470.23.14.0.3.3.0.395.3102.2-6j5.11.0....0...1c.1.64.
psy-ab..10.13.2912...0j0i131k1j0i10k1.0.OTn51M9e8MQ>.

[31] BBC News, "UK "flash crash" trader Navinder Sarao pleads guilty in US," November 10, 2016, available at <http://www.bbc.com/news/business-37932250>. Drew Harwell, "A down day on the markets? Analysts say blame the machines." *The Washington Post*, February 6, 2018, available at <https://www.washingtonpost.com/news/the-switch/wp/2018/02/06/algorithms-just-made-a-couple-crazy-trading-days-that-much-crazier/?utm_term=.83176e45214a>.

[32] Neal Boudette, "Tesla's Self-Driving System Cleared in Deadly Crash," *The New York Times*, January 19, 2017, available at <https://www.nytimes.com/2017/01/19/business/tesla-model-s-autopilot-fatal-crash.html>.

[33] Nicholas Fandos, Cecilia Kang and Mike Isaac, "House Intelligence Committee Releases Incendiary Russian Social Media Ads," *The New York Times*, November 1, 2017, available at <https://www.nytimes.com/2017/11/01/us/politics/russia-technology-facebook.html. Massimo Calabresi, "Inside Russia's Social Media War on America," *Time*, Updated May 18, 2017, accessed March 8, 2018, available at <http://time.com/4783932/inside-russia-social-media-war-america/>.

[34] Will Knight, "The Dark Secret."

[35] "War at hyperspeed–Getting to grips with military robotics: Autonomous robots and swarms will change the nature of warfare," *The Economist*, January 25, 2018, available at <https://www.economist.com/news/special-report/21735478-autonomous-robots-and-swarms-will-change-nature-warfare-getting-grips>.

[36] U.S. Joint Chiefs of Staff, *Joint Publication* 1-02: Department of Defense Dictionary of Military and Associated Terms, Department of Defense, November 8, 2010 (As Amended Through 15 February 2017), 67.

[37] Debra K. Rose, "*Only in the Mind of the Enemy:*

*Can Deterrence Effectiveness be Measured*," (Master's Thesis, Joint Advanced Warfighting School, 2011), 6, accessed March 11, 2018, available at <www.dtic.mil/dtic/tr/fulltext/u2/a545543.pdf>.

[38] Central Intelligence Agency, "*Khrushcrev–A Personality Sketch*," OCI No. 2391/61, CIA-RDP79S00427A000100020010-6, available at <https://www.cia.gov/library/readingroom/docs/CIA-RDP79S00427A000100020010-6.pdf>.

[39] Zachary Cohen, "CIA: North Korean leader Kim Jong Un isn't crazy," *CNN*, October 6, 2017, available at <https://www.cia.gov/library/readingroom/docs/CIA-RDP79S00427A000100020010-6.pdf>.

[40] Brett Mulligan, "Second Punic War (218-201 BC)," *Dickinson College Commentaries*, accessed March 11, 2018, available at <http://dcc.dickinson.edu/nepos-hannibal/second-punic-war>.

[41] Frontline, "*The Gulf War*," television broadcast, Public Broadcasting Service, January 9, 1996.

[42] James Smith and Gwendolyn Hall, "Milestones in Strategic Arms Control, 1945-2000: United States Air Force Roles and Outcomes," September 2002 (US Air Force Institute for National Security Studies, Air University Press, Maxwell AFB, AL), 90, available at <https://media.defense.gov/2017/Apr/06/2001728010/-1/-1/0/B_0087_SMITH_HALL_STRATEGIC_ARMS_CONTROL.PDF>.