# FROM ALGORITHMS TO ARMAMENTS: THE RISING ROLE OF AI IN MODERN WARFARE

Dr. Harman Shergill, Associate Professor, Lincoln College of Law, Sirhind[1]

**ABSTRACT**

Few developments in science and technology hold as much promise for the future of humanity as the suite of computer-science-enabled capabilities that falls under the umbrella of artificial intelligence (AI). While revolutionary technologies like AI hold much promise for humanity, however when used for military purposes, they can pose potential risks. The challenge is to build an understanding among stakeholders about a technology and develop responsive solutions to mitigate such risks. That is where we might be today with military applications of artificial intelligence (AI). There can be little doubt that AI has potential uses that could improve the health and well-being of individuals, communities, and states, and help meet the UN Sustainable Development Goals. However, certain uses of AI could undermine international peace and security if they raise safety concerns, accelerate conflicts, or loosen human control over the means of war. This Paper intends to provide a starting point for more robust dialogue among governments, industry and research institutions, as stakeholders endeavour to maximize the benefits of AI while mitigating the misapplication of this important technology.

**Keywords:** Artificial Intelligence, technology, military, AI

[1] Associate Professor, Lincoln College of Law, Sirhind.

## INTRODUCTION

Recent years have seen an explosion in the possibilities enabled by artificial intelligence (AI), driven by advances in data, computer processing power, and machine learning.[2] AI is disrupting a range of industries and has similar transformative potential for international relations and global security. At least two dozen countries have released national plans to capitalize on AI, and many states are seeking to incorporate AI to improve their national defense.[3] This paper aims to improve understanding of how militaries might employ AI, where those uses might introduce risks to international peace and security, and how states might mitigate these risks.[4] Artificial intelligence is not a discrete technology like a fighter jet or locomotive, but rather is a general-purpose enabling technology, like electricity, computers, or the internal combustion engine. As such, AI will have many uses. In total, these uses could lead to economic growth and disruption on the scale of another industrial revolution. This AI-driven cognitive revolution will increase productivity, reduce automobile accidents, improve health outcomes, and improve efficiency and effectiveness in a range of industries. Many, but not all, of the recent advances in AI come from the field of machine learning, in which machines learn from data, rather than follow explicit rules programmed by people.[5] AI continues to advance

---

[2] Artificial intelligence is the field of study devoted to making machines intelligent.Intelligence measures a system's ability to determine the best course of action to achieve its goals in a wide range of environments. Today's AI systems exhibit narrow artificial intelligence, or task-specific intelligence. The field of AI has a number of subdisciplines and methods used to create intelligent behavior, and one of the most prominent is machine learning. For more on definitions of artificial intelligence, see Nils J. Nilsson, The Quest for Artificial Intelligence: A History of Ideas and Achievements (Cambridge: Cambridge University Press, 2010). For more on definitions of intelligence, see Shane Legg and Marcus Hutter, A Collection of Definitions of Intelligence, technical report, Dalle Molle Institute for Artificial Intelligence, June 15, 2007, https:// arxiv.org/pdf/0706.3639.pdf. On machine learning, see Tom Michael Mitchell, "The Discipline of Machine Learning," 2006, Carnegie Mellon University, School of Computer Science, Machine Learning Department; and Ben Buchanan and Taylor Miller, Machine Learning for Policymakers: What It Is and Why It Matters, Belfer Center, June 2017, https:// www.belfercenter.org/sites/default/files/files/publication/ MachineLearningforPolicymakers.pdf. For a brief, nontechnical overview of AI and machine learning, see Paul Scharre and Michael C. Horowitz, Artificial Intelligence: What Every Policymaker Needs to Know, Center for a New American Security, June 2018, https://www.cnas.org/publications/ reports/artificial-intelligence-what-every-policymakerneeds-to-know.

[3] Tim Dutton, "An Overview of National AI Strategies," Medium. com, June 28, 2018, https://medium.com/politics-ai/ an-overview-of-national-ai-strategies-2a70ec6edfd.

[4] This paper does not consider second-order effects on international peace and security due to potential political, economic, and societal disruption from AI. These indirect effects of the AI revolution on international security are potentially even more significant, however. For more on these potential second-order effects, see Michael C. Horowitz et al., Artificial Intelligence and International Security, Center for a New American Security, July 10, 2018, https://www.cnas.org/publications/reports/ artificial-intelligence-and-international-security.

[5] For example, the AI system Pluribus, a joint project by researchers at Carnegie Mellon University and Facebook, achieved superhuman performance in no-limit Texas hold 'em poker without using any machine learning. James Vincent, "Facebook and CMU's 'Superhuman' Poker AI beats Human Pros," the Verge, July 11, 2019, https://www.theverge. com/2019/7/11/20690078/ai-poker-pluribus-facebook-cmutexas-hold-em-six-player-no-limit.

as a field of study,[6] but even if all progress were to stop today (which is unlikely),[7] there would still be many gains across society by applying current AI methods to existing problems.

The net effect of AI across society is likely to be very beneficial, but both malign and responsible actors will use AI in security applications as well. Better understanding these uses, and how to counter them when necessary, is essential to ensuring that the net effect of AI on society is maximally beneficial. State and nonstate actors have already caused harm through the deliberate malicious use of AI technology. As AI technology moves rapidly from research labs to the real world, policy makers, scholars, and engineers must better understand the potential risks from AI in order to mitigate against harm.[8]

**WAR + AI**

As a general-purpose enabling technology, AI has many potential applications to national defense. Military use of AI is likely to be as widespread as military use of computers or electricity. In the business world, technology writer Kevin Kelly has said, "There is almost nothing we can think of that cannot be made new, different, or interesting by infusing it with" greater intelligence. To imagine business applications, "Take X and add AI."[9] The same is true for military AI applications. AI is likely to affect strategy, operations, logistics, personnel, training, and every other facet of the military. There is nothing intrinsically concerning about the militarization of artificial intelligence, any more than the militarization of computers or electricity is concerning. However, some specific military applications of AI could be harmful, such as lethal autonomous weapons or the application of AI to nuclear operations. Additionally, the net effect of the "intelligentization" or "cognitization" of military operations could alter

[6] Some of the most impressive basic research advances in AI come out of a method called deep reinforcement learning, in which machines learn by interacting with the environment. This method has been used to achieve superhuman performance in complex computer games without any human training data or preprogrammed rules of behavior. For more information, see OpenAI, "OpenAI Five," https://openai.com/ five/.

[7] There are wide-ranging debates among AI researchers about the future direction of the field. For more on a few of these views, see Rich Sutton, "The Bitter Lesson," incompleteideas. net, March 13, 2019, http://www.incompleteideas.net/ IncIdeas/BitterLesson.html; and Rodney Brooks, "A Better Lesson," Robots, AI, and Other Stuff (blog), rodneybrooks.com, March 19, 2019, https://rodneybrooks.com/a-better-lesson/.

[8] For some examples of security-related applications of artificial intelligence, see Miles Brundage et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, February 2018, https://maliciousaireport.com/

[9] Kevin Kelly, "The Three Breakthroughs That Have Finally Unleashed AI on the World," Wired, October 27, 2014, https:// www.wired.com/2014/10/future-of-artificial-intelligence/.

warfare in profound ways.[10]

The first and second Industrial Revolutions dramatically changed warfare, increasing the scope and scale of destruction that could be inflicted with industrial-age weapons. Policy makers at the time were unprepared for these changes, and the result was two global wars with tens of millions of lives lost. This increased scale of destruction was not due to one or two specific uses of industrial technology in war but rather the net effect of industrialization. The Industrial Revolutions enabled the mass mobilization of entire societies for "total war," as nations turned the increased productivity and efficiency made possible by industrial technology to violent ends. Steel and the internal combustion engine made it possible to build war machines like the tank, submarine, and airplane and to take warfare to new domains under the sea and in the air. Mechanization enabled an expansion of destructive capacity through weapons like the machine gun, leading to the deadly trench warfare of World War I. And radio communications enabled coordinated long-distance operations, making possible lightning advances like the blitzkrieg of World War II.

As warfare transitioned to the Atomic Age, the extreme destructive potential of nuclear weapons was made clear in the aftermath of the bombings of Hiroshima and Nagasaki. Policy makers understood the stakes of nuclear-era warfare and the existential risk it posed—and still poses—to humanity. Yet the effect of AI on warfare is more likely to be similar to that of the Industrial Revolution, with myriad changes brought about by the widespread application of general-purpose technologies, rather than a single discrete technology like nuclear weapons.

Industrialization increased the physical scope and scale of warfare, allowing militaries to field larger, more-destructive militaries that could move farther and faster, delivering greater firepower, and in a wider array of domains. Artificial intelligence is bringing about a cognitive revolution, and the challenge is to anticipate the broad features of how this cognitive revolution may transform warfare.

[10] For an English-language analysis of Chinese military scholarship on the intelligentization of warfare, see Elsa Kania, Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power, Center for a New American Security, November 28, 2017, https://www.cnas.org/publications/reports/battlefield-singularity-artificial-intelligence-militaryrevolution-and-chinas-future-military-power.

## FEATURES OF ARTIFICIAL INTELLIGENCE

### Value of AI Systems

The field of artificial intelligence comprises many methods, but the goal is to create machines that can accomplish useful cognitive tasks.[11] Today's AI systems are narrow, meaning they are only capable of performing the specific tasks for which they have been programmed or trained. AI systems today lack the broad, flexible general intelligence that humans have that allows them to accomplish a range of tasks. While AI methods are general purpose and can be applied to solve a wide range of problems, AI systems are not able to flexibly adapt to new tasks or environments on their own. Nevertheless, there are many tasks for which AI systems can be programmed or trained to perform useful functions, including in many cases at human or even superhuman levels of performance. AI systems do not always need to reach superhuman performance to be valuable, however. In some cases, their value may derive from being cheaper, faster, or easier to use at scale relative to people.

Some of the things AI systems can do include classifying data, detecting anomalies, predicting future behavior, and optimizing tasks. Real-world examples include AI systems that:

Classify data, from song genres to medical imagery.

– Detect anomalous behavior, such as fraudulent financial transactions or computer malware.

– Predict future behavior based on past data, such as recommendation algorithms for media content or better weather predictions.

– Optimize performance of complex systems, allowing for greater efficiency in operations.

In military settings, provided there was sufficient data and the task was appropriately bounded, in principle, AI systems may be able to perform similar tasks. These could include classifying military objects, detecting anomalous behavior, predicting future adversary behavior, and optimizing the performance of military systems.

[11] Examples of different AI disciplines include neural networks, evolutionary or genetic algorithms, computational game theory, Bayesian statistics, inductive reasoning, fuzzy logic, analogical reasoning, and hand-coded expert knowledge. For more background on AI, see Scharre and Horowitz, Artificial Intelligence.

## Autonomy

Artificial intelligence can also enable autonomous systems that have greater freedom to perform tasks on their own, with less human oversight. Autonomy can allow for superhuman precision, reliability, speed, or endurance. Autonomy can also enable greater scale of operations, with fewer humans needed for large-scale operations. Autonomy can allow one person to control many systems. When embedded into physical systems, autonomy can allow vehicles with forms that might be impossible if humans were onboard, or operation in remote or dangerous locations. Autonomy enables robot snakes that can slither through pipes, underwater gliders that can stay at sea for years at a time, swarms of small expendable drones, and robots that can help clean up nuclear disasters.

## Limitations of AI Systems Today

Artificial intelligence has many advantages, but it also has many limitations.[12] Today's AI systems fall short of human intelligence in many ways and are a far cry from the Cylons, Terminators, and C-3POs of science fiction.

One of the challenges of AI systems is that the narrowness of their intelligence means that while they may perform very well in some settings, in other situations their performance can drop off dramatically. A self-driving car that is far safer than a human driver in one situation may suddenly and inexplicably drive into a concrete barrier, parked car, or semitrailer.[13] A classification algorithm that performs accurately in one situation may do poorly in another. The

---

[12] For an overview of the limitations of current narrow AI systems, see Dario Amodei et al., Concrete Problems in AI Safety, Cornell University arXiv.org, July 25, 2016, 4, https://arxiv.org/ pdf/1606.06565.pdf; Dario Amodei and Jack Clark, "Faulty Reward Functions in the Wild," OpenAI (blog), OpenAI, December 21, 2016, https://blog.openai.com/faulty-reward-functions/; and Joel Lehman et al., The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities, Cornell University arXiv. org, March 8, 2018, 6, https://arxiv. org/pdf/1803.03453.pdf.

[13] Jim Puzzanghera, "Driver in Tesla Crash Excessively on Autopilot, but Tesla Shares Some Blame, Federal Panel Finds," Los Angeles Times, September 12, 2017, http://www. latimes.com/business/la-fi-hy-tesla-autopilot-20170912- story.html; "Driver Errors, Overreliance on Automation, Lack of Safeguards, Led to Fatal Tesla Crash," National Transportation Safety Board Office of Public Affairs, press release, September 12, 2017, https://www.ntsb.gov/news/ press-releases/Pages/PR20170912.aspx; "Collision Between a Car Operating with Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida," NTSB/ HAR-17/02/PB2017-102600, National Transportation Safety Board, May 7, 2016, https://www.ntsb.gov/news/events/ Documents/2017-HWY16FH018-BMG-abstract.pdf; James Gilboy, "Officials Find Cause of Tesla Autopilot Crash into Fire Truck: Report," The Drive, May 17, 2018, http://www.thedrive. com/news/20912/cause-of-tesla-autopilot-crash-into-firetruck-cause-determined-report; "Tesla Hit Parked Police Car 'While Using Autopilot,'" BBC, May 30, 2018, https:// www.bbc.com/news/technology-44300952; and Raphael Orlove, "This Test Shows Why Tesla Autopilot Crashes Keep Happening," Jalopnik, June 13, 2018, https://jalopnik. is-test-shows-why-tesla-autopilot-crashes-keephappen-1826810902.

first version of AlphaGo, which reached superhuman performance in 2016, reportedly could not play well if the size of the game board was changed from the 19-by-19- inch board on which it was trained.[14] The narrow nature of AI systems makes their intelligence brittle—susceptible to sudden and extreme failure when pushed outside the bounds of their intended use.

Failures can manifest in a variety of ways. In some cases, the system's performance may simply degrade. For example, a facialrecognition algorithm trained on people of one skin tone may perform less accurately on people of a different skin tone.[15] In other circumstances, a failure may manifest more dramatically, such as a self-driving car that suddenly attempts to drive through an obstacle. Some failures may be obvious, while others may be more subtle and escape immediate detection but nevertheless result in suboptimal outcomes. For example, a resume-sorting AI system may have a subtle bias against certain classes of individuals.[16] Because of the opaque nature of machine learning systems, it may be difficult to understand why a system has failed, even after the fact. One complicating factor for increasingly sophisticated AI systems is that their complexity makes them less transparent to human users. This means that it can be more difficult to discern when they might fail and under what conditions. For very complex systems operating in real-world environments, there is a seemingly infinite number of possible interactions between the system's programming and its environment.[17] It is impossible to predict them all. Computer simulations can help expand the scenarios a system is evaluated against, but testers are still limited by what they can imagine, and even the best simulations will never perfectly replicate the real world. Self-driving-car companies are simulating millions of driving miles every day with computers, and still there will be situations in the real world they could not have anticipated, some of which may cause accidents.[18] AI systems are also vulnerable to a range of cognitive attacks that are analogous to cyberattacks but work at the cognitive level, exploiting vulnerabilities in how the AI system

---

[14] Bob van den Hoek, "Can AlphaGo Win Lee Sedol on a Larger Size Board? Say, 4x the Size," Quora, May 14, 2016, https:// www.quora.com/Can-AlphaGo-win-Lee-Sedol-on-a-largersize-board-Say-4x-the-size.

[15] Larry Hardesty, "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems," MIT News, February 11, 2018, http://news.mit.edu/2018/study-findsgender-skin-type-bias-artificial-intelligence-systems-0212.

[16] Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women," Reuters, October 9, 2018, https://www.reuters.com/ article/us-amazon-com-jobs-automation-insight/ amazon-scraps-secret-ai-recruiting-tool-that-showed-biasagainst-women-idUSKCN1MK08G.

[17] The number of possible interactions is not technically infinite, but it is a larger number of interactions than could be reasonably calculated.

[18] John Krafcik, "Where the Next 10 Million Miles Will Take Us," Waymo, October 10, 2018, https://medium.com/waymo/ where-the-next-10-million-miles-will-take-us-de51bebb67d3.

"thinks." Examples include poisoning the data used to train an AI system or adversarial attacks that spoof AI systems with tailored data inputs, causing them to generate incorrect outputs.[19]

All of these limitations are incredibly relevant in military environments, which are chaotic, unpredictable, and adversarial. Militaries will use AI systems, and those AI systems will break. They will suffer accidents, and they will be manipulated intentionally by adversaries. Any assessment of the role of AI in warfare must take into account the extreme brittleness of AI systems and how that will affect their performance on the battlefield.

## WAR IN THE COGNITIVE AGE

Artificial intelligence will introduce a new element to warfare: supplementing and augmenting human cognition. Machines, both physical and digital, will be able to carry out tasks on their own, at least within narrow constraints. Because today's AI systems are narrow, for the foreseeable future human intelligence remains the most advanced cognitive processing system on the planet. No AI system, or even suite of systems, can compare with the flexibility, robustness, and generality of human intelligence. This weakness of machine intelligence and strength of human intelligence is particularly important in warfare, where unpredictability and chaos are central elements. Warfare in the cognitive age will be partly a product of AI but also of human intelligence, which will remain a major feature of warfare for the foreseeable future.

Even though humans will remain involved, the introduction of artificial intelligence is likely to dramatically change warfare. AI will enable the fielding of autonomous vehicles that are smaller, stealthier, faster, more numerous, able to persist longer on the battlefield, and take greater risks.[20] Swarming systems will be valuable for a range of applications, including reconnaissance, logistics, resupply, medical evacuation, offense, and defense.

The most profound applications of AI are likely to be in information processing and command and control. Just as industrialization changed the physical aspects of warfare, artificial

---

[19] Anh Nguyen et al., "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," Computer Vision and Pattern Recognition, IEEE, 2015; James Vincent, "Twitter Taught Microsoft's AI Chatbot to Be a Racist Asshole in Less Than a Day," the Verge, May 24, 2016; and Nicolas Papernot et al., Practical Black-Box Attacks against Machine Learning, Cornell University arXiv.org, March 19, 2017, https://arxiv.org/pdf/1602.02697.pdf.

[20] Paul Scharre, Robotics on the Battlefield, Part II: The Coming Swarm, Center for a New American Security, October 15, 2014, https://s3.amazonaws.com/files.cnas.org/documents/CNAS_TheComingSwarm_Scharre.pdf?mtime=20160906082059.

intelligence will principally change the cognitive aspects of warfare. Militaries augmented with AI will be able to operate faster and with more-numerous systems, and conduct more-complex and distributed operations.

While much of the attention on military AI applications has focused on robotics, it is worth noting that in computer games, such as Dota 2, computers have achieved superhuman performance while playing with the same units as human competitors.[21] Computers' advantages have come in better and faster information processing, and command and control. Whereas humans can only pay attention to a limited number of things, an AI system can simultaneously absorb and process all incoming information at once. Machines can then process this information faster than humans and coordinate the simultaneous rapid responses of military units. These advantages will make AI systems valuable for militaries in improving battlefield awareness, command and control, and speed, precision, and coordination in action. Because of machines' limitations in responding to novel situations, however, humans will still be needed in real-world combat environments, which are more complex and unrestricted than computer games. The most effective militaries are likely to be those that optimally combine AI with human cognition in so-called centaur approaches, named after the mythical halfhuman, half-horse creature.

## POTENTIAL RISKS FROM MILITARY AI APPLICATIONS

The introduction of AI could alter warfare in ways both positive and negative. It can be tempting to envision AI technologies as principally enabling offensive operations, but they will be valuable for defensive operations as well. Because AI is a general-purpose technology, how it shifts the offense-defense balance in different areas may depend on the specific application of AI, and may evolve over time.

Some general characteristics of AI and attendant risks are outlined below, but it is worth noting that these risks are only possibilities. Technology is not destiny, and states have choices about how to use AI technology. How these risks manifest will depend on what choices states make. A concerted effort to avoid these risks may be successful.

[21] OpenAI, "OpenAI Five."

### *Accident Risk*

In principle, automation has the potential to increase precision in warfare and control over military forces, reducing civilian casualties and the potential for accidents that could lead to unintended escalation. Automation has improved safety in commercial airline autopilots and, over time, will do so for selfdriving cars. However, the challenge in achieving safe and robust self-driving cars in all weather and driving conditions points to the limitations of AI today. War is far more complex and adversarial than driving or commercial flying.

An additional problem militaries face is a lack of available data on the wartime environment. To build self-driving cars that are robust to a range of driving conditions, the autonomous car company Waymo has driven over 10 million miles on public roads. Additionally, it is computer simulating 10 million driving miles every day.[22] This allows Waymo to test its cars under a variety of conditions. The problem for militaries is that they have little to no ground-truth data about wartime conditions on which to evaluate their systems. Militaries can test their AI systems in training environments, either in the real world or in digital simulations, but they cannot test their actual performance under real operational conditions until wartime. Wars are a rare occurrence, fortunately. This poses a problem for testing autonomous systems, however. Militaries can do their best to mimic real operational conditions as closely as possible in peacetime, but they can never fully recreate the chaos and violence of war. Humans are adaptable and are expected to innovate in wartime, using their training as a foundation. But machine intelligence is not as flexible and adaptable as human intelligence. There is a risk that military AI systems will perform well in training environments but fail in wartime because the environment or operational context is different, perhaps even only slightly different. Failures could result in accidents or simply cause military systems to be ineffective.

Accidents with military systems could cause grave damage. They could kill civilians or cause unintended escalation in a conflict. Even if humans regained control, an incident that killed adversary troops could escalate tensions and inflame public sentiment such that it was difficult for national leaders to back down from a crisis. Accidents, along with vulnerabilities to hacking, could undermine crisis stability and complicate escalation management among nations.

[22] Krafcik, "Where the Next 10 Million Miles Will Take Us."

### *Autonomy and Predelegated Authority*

Even if AI systems perform flawlessly, one challenge nations could face is the inability to predict themselves what actions they might want to take in a crisis. When deploying autonomous systems, humans are predelegating authority for certain actions to a machine. The problem is that in an actual crisis situation, leaders may decide that they want to take a different approach. During the Cuban Missile Crisis, US leaders decided that if the Soviets shot down a US reconnaissance plane over Cuba, they would attack. After the plane was shot down, they changed their minds. Projection bias is a cognitive tendency where humans fail to accurately predict their own preferences in future situations.

The risk is that autonomous systems perform as programmed, but not in ways that human leaders desire, raising the risk of escalation in crises or conflicts.

### *Prediction and Overtrust in Automation*

Maintaining humans in the loop and restricting AI systems to only giving advice is no panacea for these risks. Humans frequently overtrust in machines, a phenomenon known as automation bias.[23] Humans were in the loop for two fratricide incidents with the highly automated US Patriot air and missile defense system in 2003 yet failed to stop the accidents.[24] In one notable psychological experiment, participants followed a robot the wrong way through a smoke-filled building that was simulating a fire emergency, even after being told the robot was broken.[25]

Overtrusting in machines could lead to accidents and miscalculation, even before a war begins. In the 1980s, the Soviet Union conducted Operation RYaN to warn of a surprise US nuclear attack. The intelligence program tracked data on various potential indicators of an attack, such as the level of blood in blood banks, the location of nuclear weapons and key decisionmakers, and the activities of national leaders.[26] If AI systems could actually give accurate early warning of a surprise attack, this could be stabilizing. Knowing that there was no possibility of

---

[23] Kate Goddard et al., "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," Journal of the American Medical Informatics Association 19, no. 1 (2012): 121-7, https://www.ncbi.nlm.nih.gov/pmc/articles/ PMC3240751/.

[24] Paul Scharre, Army of None: Autonomous Weapons and the Future of War (New York: W. W. Norton, 2018), 137-145.

[25] Paul Robinette et al., Overtrust of Robots in Emergency Evacuation Scenarios, 2016, https://www.cc.gatech. edu/~alanwags/pubs/Robinette-HRI-2016.pdf.

[26] Bernd Schaefer et al., Forecasting Nuclear War: Stasi/ KGB Intelligence Cooperation under Project RYaN, Wilson Center, November 13, 2014, https://www.wilsoncenter.org/ publication/forecasting-nuclear-war.

successfully carrying out a surprise attack, nations might refrain from attempting one. Yet prediction algorithms are only as good as the data on which they are trained. For rare events like a surprise attack, there simply isn't enough data available to know what is actually indicative of an attack. Flawed data will lead to flawed analysis. Yet the black-box nature of AI, in which its internal reasoning is opaque to human users, can mask these problems. Without sufficient transparency to understand how the algorithm functions, human users may not be able to see that its analysis has gone awry.

### *Nuclear Stability Risks*

All of these risks are especially consequential in the case of nuclear weapons, where accidents, predelegated authority, or overtrust in automation could have grave consequences. False alarms in nuclear early warning systems, for example, could lead to disaster. There have been numerous nuclear false alarms and safety lapses with nuclear weapons throughout the Cold War and afterward.[27] In one particularly notable incident in 1983, a Soviet early warning satellite system called Oko falsely detected a launch of five US intercontinental ballistic missiles against the Soviet Union. In fact, the satellites were sensing the reflection of sunlight off of cloud tops, but the automated system told human operators "missile launch." Soviet Lieutenant Colonel Stanislav Petrov judged the system was malfunctioning, but in future false alarms, the complexity and opacity of AI systems could lead human operators to overtrust those systems.[28] The use of AI or automation in other aspects of nuclear operations could pose risks as well. For example, nuclear-armed uninhabited aircraft (drones) could suffer accidents, leading states to lose control of the nuclear payload or accidentally signaling *escalation to an adversary.*

### *Competitive Dynamics and Security Dilemmas*

Competition exacerbates many of these risks. Despite media headlines warning of an AI arms race, the current situation among states does not resemble previous arms races, in which countries spent escalating sums of money on battleships or nuclear weapons without gaining

---

[27] Patricia Lewis et al., Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy, Royal Institute of International Affairs, London, April 2014, https://www. chathamhouse.org/publications/papers/view/199200; and Scott D. Sagan, The Limits of Safety: Organizations, Accidents, and Nuclear Weapons (Princeton, NJ: Princeton University Press, 1993)

[28] David Hoffman, "'I Had a Funny Feeling in My Gut,'" Washington Post, February 10, 1999, http://www. washingtonpost.com/wp-srv/inatl/longterm/coldwar/ shatter021099b.htm.

any clear military advantage. AI innovation today is largely driven by the commercial sector, and militaries seek to import AI technology to defense applications. Competitive dynamics could still lead to security dilemmas, in which states individually take actions to increase their own security, but with the net effect of decreasing security for all. The two greatest risks in a race to use AI are in speed and safety.

### *Speed*

One of the great dangers of automation is an arms race in speed, in which countries push humans further and further out of the loop in a bid to act faster than competitors. The consequences of this dynamic can be seen in stock trading, which is highly automated today. Algorithms execute trades at speeds measured in microseconds (1 microsecond equals 0.000001 seconds).[29] In a single eyeblink, 100,000 microseconds pass by. Yet when algorithms get it wrong, they can wreak havoc at machine speed. In the May 2010 "flash crash," a combination of brittle algorithms, highfrequency trading, market instability, and humans taking advantage of predictable bot behavior all combined to create a perfect storm in which the US stock market lost nearly 10 percent of its value in minutes.[30] Two years later, the high-frequency trading firm Knight Capital Group suffered an accident with a runaway algorithm, which began making erroneous trades at machine speed, moving $2.6 million a second. Within 45 minutes, it had lost $460 million, more than the company's entire assets.[31]

Financial regulators have dealt with the problem of flash crashes not by preventing them from occurring but by installing circuit breakers that take stocks offline if prices move too quickly and mitigate the consequences of an event.[32] Miniflash crashes continue to occur, and in a single day in 2015, over 1,200 circuit breakers were tripped across multiple financial markets around the globe.[33]

---

[29] Michael Lewis, Flash Boys: A Wall Street Revolt (New York: W. W. Norton, 2015), 63, 69, 74, 81.

[30] US Commodity Futures Trading Commission and US Securities and Exchange Commission, "Findings Regarding the Market Events of May 6, 2010," September 30, 2010, 2, http://www.sec.gov/news/studies/2010/marketeventsreport.pdf.

[31] D7, "Knightmare: A DevOps Cautionary Tale," Doug Seven (blog), April 17, 2014, https://dougseven.com/2014/04/17/ knightmare-a-devops-cautionary-tale/.

[32] US Securities and Exchange Commission, "Investor Bulletin: Measures to Address Market Volatility," July 1, 2012, https:// www.sec.gov/oiea/investor-alerts-bulletins/investor-alertscircuitbreakersbulletinhtm.html.

[33] Matt Egan, "Trading Was Halted 1,200 Times Monday," CNN Money, August 24, 2015, http://money.cnn.com/2015/08/24/ investing/stocks-markets-selloff-circuit-breakers-1200- times/index.html.

An escalatory incident between competitive military AI systems could have serious consequences. The challenge nations face is that there are no referees to call time out in war. If nations are to prevent such an incident, they will need to build in their own circuit breakers to limit the potential consequences of automation. These risks are particularly acute in cyberspace, where cyber systems could have global effects in seconds. A flash war would benefit no one.

Even once a war begins, an AI-accelerated operational tempo could lead to less human control over battlefield operations. Some Chinese scholars have hypothesized about a "battlefield singularity" in which the pace of action eclipses human decision making, and some US scholars have used the term "hyperwar" to refer to a similar situation.[34] The problem is that greater speed on one side necessitates greater speed on the other, with a net outcome that is more harmful for all. Moving to a new domain of warfare with less human control would be dangerous and risk large-scale accidents or escalation, even within a conflict. All militaries have an incentive to keep war more effectively under human control.

***Race to the Bottom on Safety***

Speed is not only a concern on the battlefield but also in peacetime development and deployment of military systems. Testing and evaluation are vitally important for improving the safety of complex autonomous systems. Greater testing in real-world and simulated environments can help identify flaws in a system ahead of time and reduce the risk of accidents. While no amount of testing can render a system 100 percent accident proof, moreextensive testing can help reduce the risk of accidents.

Unfortunately, a desire to beat a competitor to fielding a new system could cause actors to cut corners on safety, deploying autonomous systems before they are ready. This speed-to-market dynamic has been implicated as a possible contributing factor to accidents in the commercial airline autopilot industry and self-driving cars. If such a dynamic were to befall militaries, the

---

[34] Chen Hanghui [陈航辉], "Artificial Intelligence: Disruptively Changing the Rules of the Game" [人工智能：颠覆性改变"游戏规则], China Military Online, March 18, 2016, http:// www.81.cn/jskj/2016-03/18/content_6966873_2.htm (Chen Hanghui is affiliated with the Nanjing Army Command College); and John R. Allen and Amir Husain, "On Hyperwar," Proceedings, July 2017, https://www.usni.org/magazines/proceedings/2017/july/hyperwar.

result would be a world of unreliable military AI systems, which would make all nations less safe.[35]

## MITIGATING POTENTIAL RISKS

Nations build militaries precisely because they don't trust others and want to provide for their own defense. In spite of this, states have come together on many occasions to limit the proliferation, development, production, stockpiling, or use of various military technologies that were seen as excessively harmful, inhumane, or destabilizing. Arms control is one option for mitigating risks from AI, but there are other unilateral measures states can take.

### *Technology Controls*

Military technologies can be controlled or restricted at a number of stages along their development cycle. Nonproliferation regimes aim to limit access to the underlying technology behind certain weapons. The Nuclear Non-Proliferation Treaty, for example, aims to prevent the spread of nuclear weapons, promote cooperation on peaceful uses of nuclear energy, and further the goal of nuclear disarmament. Some weapons bans, like those on land mines and cluster munitions, allow access to the technology but prohibit developing, producing, or stockpiling the weapon. Other bans only apply to use, sometimes prohibiting use entirely or proscribing only certain kinds of uses of a weapon. Finally, arms-limitation treaties permit use but limit the quantities of certain weapons states can have in peacetime.[36]

AI is not like nuclear technology; it is more like computers, which are diffuse and driven by the commercial sector.[37] AI research papers are openly published online, and trained AI models can often be downloaded for free from online resources. Many actors will have access to AI, and preventing the underlying availability of AI is not likely to be feasible, at least given the shape of AI technology today. However, the specific uses of AI are more important, and states

---

35 For more on the risk of a race to the bottom on safety, see Paul Scharre, "Killer Apps: The Real Dangers of an AI Arms Race," Foreign Affairs, (May/June 2019), https://www. foreignaffairs.com/articles/2019-04-16/killer-apps.

36 For a comprehensive overview of different types of weapons bans, see Scharre, Army of None, 331-345.

37 For controls on other dual-use technologies, see Elisa D. Harris, ed., Governance of Dual-Use Technologies: Theory and Practice (Cambridge, MA: American Academy of Arts and Sciences, 2016), http://www.amacad.org/sites/default/files/academy/multimedia/pdfs/publications/researchpapersmonographs/GNF_Dual-Use-Technology. pdf; and Jonathan B. Tucker, ed., Double-Edged Innovations: Preventing the Misuse of Emerging Biological/Chemical Technologies, Defense Threat Reduction Agency, July 2010, https://apps.dtic.mil/dtic/tr/fulltext/u2/a556984.pdf.

have choices about how the technology is used. Bans on land mines and cluster munitions don't prohibit access to the technology, but they do prohibit producing, stockpiling, or using those weapons. Arms control over AI as a whole would likely be infeasible, like attempting arms control for industrialization.

However, the Industrial Revolution saw a raft of treaties on various applications of industrial technology to war, treaties that had a mixed track record of success in the late 19th and early 20th centuries. Similarly, it is possible to conceive that arms control on some applications of AI could be successful. Achieving trust among all parties would be challenging, since AI systems are software and not observable in the same way naval ships or nuclear missiles are, which permits states to verify that others are complying with the treaty. However, there may be ways to achieve sufficient verification and compliance through other means or on some aspects of AI.

Transparency and confidence-building measures could also help reduce the risk of accidents by reducing the potential for miscalculation or misunderstanding among states.[38]

### *Building Safe and Secure AI Systems*

Ultimately, the most powerful tool states have at their disposal for mitigating the risk of military AI systems comes from building safe and secure AI systems themselves. Militaries have an incentive to keep their systems under effective operational control. AI systems that slip out of human control could not only cause an accident, possibly harming third parties, but are also not very useful to the military that deploys them. Military systems that may not work or could be hacked by the enemy are not very useful or valuable. Conducting better tests and evaluation and maintaining humans in overall operational control of the system through a humanmachine centaur command-and-control model may be the best approach for mitigating the risks of military AI.

[38] For example, see United Nations, "Group of Governmental Experts on Transparency and Confidence-Building Measures in Outer Space Activities," July 29, 2013, https://undocs. org/A/68/189.