The AI-cyber nexus: implications for military escalation, deterrence and strategic stability

Author: James Johnson, Postdoctoral fellow, James Martin Center for Nonproliferation Studies (CNS), Middlebury Institute of International Studies, Monterey

Abstract: How could AI-infused cyber capabilities be used to subvert, or otherwise compromise, the reliability, control and use of states' nuclear forces? This article argues that a new generation of artificial intelligence (AI) enhanced cyber capabilities will amplify the risk of inadvertent escalation caused by the co-mingling of nuclear and strategic non-nuclear weapons and the increasing speed of warfare, thereby increasing the risk of nuclear confrontation. It examines the potential implications of cyber (offensive and defensive) capabilities augmented with AI applications for nuclear security. The article concludes that future iterations of AI-enhanced cyber counterforce capabilities will complicate the existing challenges of cyber defence, and in turn, compromise nuclear assets and increase the escalatory effects of offensive cyber capabilities.

Keywords: Artificial Intelligence; Cyber-security; US-China Relations; Nuclear Security; Deterrence Policy; Emerging Technology

Introduction

Given the hyperbole surrounding AI, it is easy to overstate the opportunities and challenges posed by the development and deployment of AI in the military sphere (Boulanin 2019). Today, there remains a large amount of debate among AI researchers surrounding the significance of several significant technical and operational challenges in the deployment of AI-infused systems (e.g. drone swarming, command and control decision-making support systems and a broader range of autonomous weapon systems). Moreover, many of the risks posed by AI in the nuclear domain today are not necessarily new (Boulanin 2019). Recent advances in AI (especially machine learning techniques) exacerbate existing risks to escalation and stability rather than generating entirely new ones. While AI could enable significant improvements in many military domains (including nuclear weapons), for the foreseeable future, developments in military AI will likely be far more prosaic than implied in popular culture. The main worry for nuclear stability, therefore, is that militaries will underestimate (or overstate) or ignore the potential shortcomings and risks associated with the use of the current generation of AI technology in the safety-critical military sphere – and especially the nuclear domain.

Because of the diverse approaches to AI-based research, there is no universally accepted definition of 'artificial intelligence,' which is confusing when the term is applied generically to make grandiose claims about its revolutionary effects.[1] AI can be understood as a universal term for improving the performance of automated (or autonomous) systems to solve a wide variety of complex tasks including: *perception* (sensors, computer vision, audio and image processing); *reasoning and decision-making* (problem solving, searching, planning and reasoning); *learning and knowledge representation* (machine learning, deep networks and modelling);[2] *communication* (language processing); *autonomy and robotics*; and *human-AI collaboration* (humans define the systems' purpose, goals and context).

In a military context, as a potential enabler and force multiplier of advanced weapon systems, AI is more akin to electricity, radio, radar and intelligence, surveillance and reconnaissance (ISR) support systems than a 'weapon' *per se*. Much like these enabling support systems, AI is also strategically vital, vulnerable, inherently cross-domain and dependent on advanced technology. Thus, even if AI applications are unable to make better battlefield decisions than humans,[3] militaries that use AI in human-machine teaming will doubtless gain significant advantages (e.g. remote-sensing, situational-awareness, battlefield-manoeuvres and a compressed decision-making loop), compared to those who depend on human judgment – and semi-autonomous technology – alone. This is particularly the case, in operating systems in operating environments that demand endurance and rapid decision-making across multiple combat zones (Ayoub and Payne 2016).

[1] Historical definitions of AI can be grouped as follows: systems that think like humans; systems that think rationally; systems that act like humans; and systems that function rationally.

[2] Applied in a non-technical context, the term 'learning' can sometimes be a cause of confusion, because it can presume an anthropomorphic interpretation.

[3] In contrast to human decision makers, cognitive stressors, time pressures and other physical effects of combat (such as lack of glucose and fatigue), do not adversely affect AI systems (Ayoub and Payne 2016).

This article argues that a new generation of AI-augmented offensive cyber capabilities will amplify the risk of inadvertent escalation posed by the co-mingling of nuclear and strategic (or counterforce) non-nuclear weapons and the increasing speed of warfare, and in turn, increase the risk of nuclear confrontation. The article's thesis is grounded in three core themes. First, AI does not exist a vacuum, that is, in isolation, AI will unlikely be a strategic game changer. Instead, it will likely mutually reinforce the destabilizing effects of existing advanced capabilities thereby increasing the speed of warfare and compressing the decision-making timeframe (Johnson, 2019a). AI-enabled and enhanced capabilities will have a more significant impact (positive or negative) on strategic stability than the sum of its parts. Furthermore, the intersection of AI with nuclear weapons, and a broader spectrum of strategic non-nuclear weapons, will likely accelerate the erosion of the survivability of nuclear arsenals associated with the 'computer revolution.' Put another way, military AI and the advanced capabilities it enables are a natural manifestation – rather than the cause or origin – of an established trend, which could lead states to adopt destabilizing launch postures due to the increasing speed of war and co-mingling.

Second, AI's impact on stability, deterrence and escalation will likely be determined as much (or more) by states' *perception* of its functionality as much as by what it is capable of doing. In the case of nuclear strategy and deterrence, the perception of an adversary's capabilities and intentions is as (or more) important than its *actual capability*. Thus, in addition to the importance of military force postures, capabilities and doctrine, the effects of AI will also have a strong cognitive element, increasing the risk of inadvertent escalation as a result of misperception and misunderstanding.

Finally, and related to the above, the increasingly competitive and contested nuclear multipolar world order will compound the destabilizing effects of AI, and in turn, increase the escalation risks in future warfare between great military powers – especially China and the United States (Johnson 2019b). Moreover, the potential operational and strategic advantages offered by AI-augmented capabilities could prove irresistible to nuclear-armed strategic rivals, causing them to eschew the limitations of AI, and compromise safety and verification standards to protect, or attempt to capture, technological superiority on the future digitized battlefield.

Military AI primer

Artificial intelligence research began as early as the 1950s, as a broad concept concerned with the science and engineering of making intelligent machines. Since the 1950s, AI research has gone through several development phases – from early exploitations in the 1950s and 1960s and the 'AI Summer' during the 1970s, through to the early 1980s and the 'AI Winter' from the 1980s – each of which failed to live up to its initial, and often over-hyped, expectations. In the past decade, the explosion of interest in the field (or 'AI renaissance') occurred due to the convergence of four enabling developments: (1) the exponential growth in computing processing power; (2) expanded data sets (especially 'big data' sources) (Gray et al. 2015)[4]; (3) advances in the implementation of machine learning techniques and algorithms (especially deep 'neural networks') (Schmidhuber 2015);[5] and (4) the rapid expansion of commercial interest and investment in AI (Hoadley and Lucas 2018).[6]

AI is concerned with machines that emulate capabilities that are usually associated with human intelligence, such as language, reasoning, learning, heuristics and observation. Today, all practical (i.e. technically feasible) AI applications fall into the 'narrow' category, and less so, artificial general intelligence (AGI) or 'superintelligence' (Reedy 2017)[7]. 'Narrow AI' has been widely used in a broad range of civilian and military tasks since the 1960s, (Russel and Norvig 2014; Nilsson 2010) and involves statistical algorithms (mostly based on machine learning techniques) that learn procedures through analysis of large training data sets designed to approximate and replicate human cognitive tasks (Russel and Norvig 2014)[8]. 'Narrow AI' is the

[4] Though many AI systems rely on large amounts of data, AI does not necessarily entail the volume, velocity and variety usually associated with 'Big Data' analytics (Gray et al. 2015).

[5] 'Machine learning' is a concept that encompasses a wide variety of techniques designed to identify patterns in, and 'learn' and make predictions from, data sets. Successful 'learning' depends on having access to vast pools of reliable data about past behaviour and successful outcomes (Schmidhuber 2015).

[6] US technology companies had reportedly invested an estimated $20-$30 billion in narrow AI algorithms in 2016, and this amount is expected to reach $126 billion by 2025. (Hoadley and Lucas 2018).

[7] Recent progress in AI falls within two distinct fields: (1) 'narrow' AI, and particularly, machine learning; (2) 'general' AI, which refers to AI with the scale and fluidity akin to the human brain. Most AI researchers anticipate 'general' AI to be at least several decades away. Narrow AI is already utilized in the private sector, in particular, in data-rich research fields and applied sciences. Most experts generally agree that the development of 'general' AI is at least several decades away, if possible, at all (Reedy 2017).

[8] During the 'learning' process, machine learning algorithms generate statistical models to accomplish a specified task in situations it has not previously encountered (Russel and Norvig 2014).

category of artificial intelligence to which this article refers when it assesses the impact of this technology in a military context.

Given the diverse approaches to research in AI (Boden 2016; Vernon 2014)[9], there is no universally accepted definition of AI. A recent US congressional report defines AI as follows: ‘Any artificial system that performs tasks under varying and unpredictable circumstances, without significant human oversight, or that can learn from their experience and improve their performance...they may solve tasks requiring human-like perception, cognition, planning, learning, communication, or physical action’ (Hoadley and Lucas 2018).

Today, a large gap exists between the reality of what AI technology is capable of doing in a military context, and the expectations and fears of public opinion, policymakers and global defence communities. How will AI technologies influence public opinion about the use of military force and warfare? (Todd, Narang, and Talmadge 2019). The misrepresentations and misconceptions that exist today in the narratives surrounding AI (especially in societal, economic and national security debates) and are in large part caused by the hyperbole and exaggerated depictions of AI in science fiction (Zarkadakis 2015). [10] Misrepresentations of the potential opportunities and risks associated with this technology in the military sphere can obscure constructive and crucial debate on these topics. Specifically, this includes the challenge of balancing the potential operational, tactical and strategic benefits of leveraging AI while managing the risks posed to nuclear security in the (especially premature) pursuit of these advantages.

The historical record on technological change in a military context counsels scepticism against alarmist predictions and extrapolations from emerging trends: technologies rarely evolve in the way futurists predict and many applications have had countervailing or conditional effects that have ameliorated the dystopian predictions of naysayers (Mets 2000; Mets and Kievit 1995). In short, AI’s impact on nuclear stability will, like previous generations of emerging technology (i.e. chemical and biological weapons, cyber and nuclear weapons), have both stabilizing and destabilizing (and possibly contradictory) effects on strategic stability. For example, many expected that chemical weapons would instantly and dramatically change the nature of warfare and

[9] AI has generally been viewed as a subfield of computer science, focused on solving computationally hard problems through search, heuristics and probability (Boden 2016; Vernon 2014).

[10] For example, see Zarkadakis (2015).

deterrence after the British used poison gas during World War I (Todd, Narang, and Talmadge 2019). However, chemical weapons proved far less practical, impactful, disruptive and relatively easier to defend against than conventional explosives. As a corollary, the degree to which military AI poses risks to future strategic stability will depend in large part on the pace and scope with which this technology facilitates new ways to improve the delivery of, and defence against, nuclear weapons and strategic non-nuclear weapons.

Conceptually, AI-augmented applications can be categorized into those that have predominately operational, tactical and strategic implications in future warfare. At the operational and tactical level, applications include: autonomy and robotics; multi-actor interaction red teaming war-gaming; [11] big data-driven modelling; [12] intelligence collection and analysis (e.g. to locate and monitor mobile missiles, troops movement) (Conklin 2018); [13] cybersecurity talent management; predictive maintenance, logistics, planning and forecasting; and vendor contract and budget management. In preparation for a new AI strategy, US Strategic Operations Command (SOCOM) recently crafted a roadmap that invests heavily in AI and machine learning technology based on several pilot projects with three overarching goals to create an algorithmic warfare multi-operational team: 'AI-ready workforce, AI-applications, and AI-outreach' (Lee 2019).

Therefore, AI's impact on strategic stability and escalation will also be shaped by broader and more nuanced factors which influence the trajectory of emerging technologies, including: military strategy and doctrine – that seeks to manipulate escalation risks – military culture and organization, alliance structures, domestic politics and public opinion, to name but a few. The Cold War record demonstrates that emerging technologies act primarily to enable independent variables which, in combination with other endogenous factors, can heighten escalation risks between adversaries. Therefore, in isolation, technology is not a major exogenous cause of (inadvertent or intentional) military escalation.

At a strategic level of warfare, AI uses include: qualitative improvements to the

[11] For example, in 2017, China's Institute of Command and Control held the nation's first-ever 'Artificial Intelligence and War-Gaming National Finals'.

[12] For example, AI is enabling scientists to model nuclear effects to confirm the reliability of the nuclear stockpile without nuclear testing.

[13] For example, the US National Geospatia-Intelligence Agency (NGA) has reportedly used AI to support military and intelligence analysis. (Conklin 2018).

nuclear command, control, communications, and intelligence (C3I) architecture; enhancing target acquisition, tracking, guidance systems and discrimination of missile and air defence systems; being force multipliers of *both* offensive and defensive machine-learning infused cyber capabilities; and qualitatively bolstering nuclear and non-nuclear missile delivery systems – including hypersonic variants.[14]

In combination, AI machine learning algorithms fused with advances in sensor technology could be engineered to allow nuclear delivery systems to operate more autonomously and precisely, with less reliance on human operators calibrating navigation and guidance parameters and equipped with more robust countermeasures against jamming or spoofing attacks. At the strategic level, AI-augmented command and control systems could potentially mitigate many of the shortcomings inherent to human strategic decision-making during the 'fog of war', such as the susceptibility to invest in sunk costs, skewed risk judgment, cognitive heuristics and groupthink.

In sum, the fusion of AI machine learning and human judgment to gauge an adversary's intentions (and predict escalation) for the purposes of planning and directing future wars for the pursuit of political objectives, is, therefore, a far less unlikely prospect in the near future than the use of AI to achieve tactical and operational ends (e.g. drone swarming and cyber defence).

Blurring the AI-cyber offence-defence line: (Glaser and Kaufmann 1998; Lynn-Jones 1995)[15]

Several US national security officials believe that AI, used as force multipliers for *both* defensive and offensive cyber weapons, will have a transformative impact on cybersecurity (Gartzke and Lindsay 2015; Wyatt 2019; Slayton 2017; Rid 2012).[16] Director of US National Intelligence, Daniel Coats, recently warned that AI could increase US vulnerability to cyberattacks, weaken its ability to attribute such attacks, improve the effectiveness and capabilities of adversaries weapon and intelligence

[14] Several states – notably China and Russia – are conducting research on the use of machine learning to develop control systems for hypersonic vehicles. Because of their high velocity, these capabilities cannot be operated manually.

[15] The offence-defence balance is derived from a broader security dilemma theory that posits states may find themselves in conflict through their attempts to increase security (i.e. acquiring weapons and offensive strategies) because they inadvertently make another state feel less secure. See Glaser and Kaufmann (1998) and Lynn-Jones (1995).

[16] Most defence analysts agree that cyberwarfare is 'offensive-dominant' in nature. See Gartzke and Lindsay (2015), Hoffman (2019) and Slayton (2017). For an opposing view, see Rid (2012).

systems, and cause accidents and related liability issues. To be sure, the line between AI cyber-offence and cyber-defence will likely remain an obscure one.[17] Bernard Brodie's words of caution about the advent of nuclear weapons almost six decades ago resonate: 'The [military] bias towards the offensive creates special problems in any *technologically new situation* where there is little or no relevant war experience to help one to reach a balanced judgment' (Brodie 1959).

On the one hand, AI might reduce a military's vulnerability to cyberattacks. AI cyber-defence tools (or 'counter-AI'), designed to recognize changes to patterns of behaviour in a network and detect anomalies, automatically detect software code vulnerabilities, and apply machine learning techniques (such as 'deep learning') to detect deviations and anomalies from normal network activity,[18] could form a more robust defence against cyber subversions (Johnson, 2019a).[19] According to the Pentagon's new AI strategic report, 'AI can enhance our ability to predict, identify, and respond to cyber and physical threats from a range of sources' (US Department of Defense, 2019). In addition, the DoD's Defense Innovation Unit (DIU) is prototyping an application (related to Project VOLTRON) that leverages AI to decipher high-level strategic questions, map probabilistic chains of events and develop alternative strategies in order to make DoD systems more resilient to AI-augmented cyberattacks and configure and fix errors more quickly than humans.

On the other hand, autonomy itself might increase a military's vulnerability to cyberattacks, which rely on stealth, deception and stratagem. For example, an adversary could use malware to take control, manipulate or fool the behaviour and pattern recognition systems of autonomous systems, such as DoD's Project Maven. Offensive attacks such as this would be relatively easy to execute, but very difficult to detect, attribute or effectively counter (Shachtman 2011).[20] This problem set is compounded by the lack of an agreed framework or understanding of what constitutes escalatory behaviour (or 'firebreaks') in cyberspace. Thus, a cyber operation intended as a signal

[17] DARPA's 2016 Cyber Grand Challenge demonstrated a single AI algorithm capable of simultaneously playing *both* offensive and defensive roles.

[18] 'Deep learning' is a type of representation learning, which in turn is a type of machine learning. This technique can enhance the ability of machines to extract salient features from a landscape or image, which could be used for classification and pattern recognition.

[19] To date, research in the field of 'counter-AI' is still at a very nascent stage. Analysts have made some progress in detecting anomalies in network behaviour, to isolate possible exploitable vulnerabilities within machine learning AI software.

[20] For example, the difficulty analysts faced in detecting malware that infected a UAV's control system at the Creech US Air Force Base in Nevada (Shachtman 2011).

(i.e. for coercive diplomacy) could go undetected by the target, or worse, misinterpreted as an offensive attack. Even if information relating to an operation of this kind is accurately identified in a timely manner, the motives behind them could remain ambiguous, or misperceived. According to Robert Jervis, 'it is likely that the country that is the object of the attack would assume that any effect was the intended one' (Jervis 2016).

US Cyber Fleet Command Commander, Michael Gilday, recently told the Senate Armed Services Committee that the US Navy must 'improve an ability to proactively *detect new and unknown malware*…so we [the US] can act quickly using advanced analytics enabled by AI and machine learning', which may give the US a 'tactical advantage' to identify malicious activity early on (Osborn 2018). Even if analysts can obtain high-quality and reliable intelligence, however, they may not want to reveal it, because doing so could compromise a source, capability or tactic (Johnson, 2019a). Moreover, most observers now acknowledge that no strategy (i.e. combining offence and defence cyber operations) could realistically be expected to deter *all* malign cyber-attacks (Hoffman 2019).[21] While AI-enhanced cyber capabilities can enhance deterrence, they can simultaneously incentivize others to attack thereby exacerbating the paradox of *enhanced capabilities* and *increased vulnerabilities* in the cyber domain (Slayton 2017; Gartzke and Lindsay 2017; Libicki 2016).[22] As the historical record attests, this capability-vulnerability paradox is heightened when states are both dependent on a particular capability (such as AI and cyber tools) and their access or ability to use the capabilities is vulnerable to an adversary's exploitation or subversion – creating first-mover incentives (Rid 2013).[23]

Cybersecurity and nuclear weapon systems: nuclear risk redux?

At a strategic level of conflict, AI applications designed to enhance cybersecurity for nuclear forces could simultaneously make cyber-dependent nuclear weapon systems

[21] Unlike nuclear deterrence, cyber deterrence does not necessarily fail if an attack is launched (Hoffman 2019).

[22] Several scholars argue that the use of offensive cyber capabilities at a strategic level of conflict is either operationally difficult or subject to successful deterrence from adversaries. Consequently, the cyber offence is considered most effective in the 'grey zone' or low-level intensity use of force. See Slayton (2017), (Lindsay and Gartzke 2017) and (Libicki 2016).

[23] A counterargument posits that fragmented and diffused information vulnerabilities associated with cyberspace, and digital networks more broadly, are less likely to cause destabilizing incentives like first-strike. See Rid (2013).

(i.e. communications, data processing or early-warning sensors) more vulnerable to cyberattacks. It is now thought possible that a cyberattack (i.e. spoofing, hacking, manipulation and digital jamming) could infiltrate a nuclear weapons system, threaten the integrity of its communications, and ultimately (and possibly unbeknown to its target) gain control of its (nuclear and non-nuclear) command and control systems.[24] AI technology has not yet evolved to a point where it would allow nuclear-armed states to credibly threaten the survivability of each other's nuclear second-strike capability. Thus, the development trajectory of AI (and its enabling technologies) means its impact on nuclear security will likely be theoretical and speculative for the foreseeable future.

Because of the intense time pressures that would likely loom large with the decision to use nuclear weapons – especially where a state maintains a launch-on-warning posture – AI-enhanced cyberattacks against nuclear systems would be almost impossible to detect and the warning signals difficult to authenticate, let alone attribute, within the short timeframe for initiating a nuclear strike. According to open sources, operators at the North American Aerospace Defence Command (NORAD) have less than three minutes to assess and confirm initial indications from early-warning systems of an incoming attack. This compressed decision-making time frame could put political leaders under intense pressure to make a decision to escalate during a crisis, with incomplete (and possibly false) information of a situation. Ironically, new technologies designed to enhance information (i.e. modernized nuclear C3I systems augmented by 5G networks, machine learning, big-data analytics and quantum computing) can also undermine clear and reliable information flow and communication, critical for effective deterrence (Gartzke and Lindsay 2019).[25]

Advances in AI could also exacerbate this cybersecurity challenge by enabling improvements to the cyber offence. Machine learning and AI by automating advanced persistent threat (APT) operations might dramatically reduce the extensive manpower resources and high levels of technical skill required to execute APT operations (or 'hunting for weaknesses'), especially against hardened nuclear targets.[26] The machine speed of AI-augmented cyber tools could enable a low-skilled and capital-restricted

[24] Recent reports of successful cyberattacks against dual-use early-warning systems suggest these claims are credible.

[25] Effective deterrence depends on the clear communication of credible threats (and consequence of violation) between adversaries. See Lindsay and Gartzke (2019).

[26] The cost of tools used to create malicious documents depends a lot on whether the malware can remain within the systems and escape detection by antivirus software.

attacker to exploit a narrow window of opportunity to penetrate an adversary's cyber-defences or use APT tools to find new vulnerabilities. For example, when docked for maintenance air-gapped nuclear-powered submarines, considered secure when submerged, could become increasingly vulnerable to a new generation of low-cost – possibly black-market – and highly automated APT cyberattacks.

An attacker could also apply AI machine learning techniques to target autonomous dual-use early-warning and other operating systems (e.g. C3I, ISR, early-warning and robotic control networks) with 'weaponized software' such as hacking, subverting, spoofing or tricking, causing unpredictable and potentially undetectable errors, malfunctions and behavioural manipulation to weapons systems – or 'data-poisoning'[27]. Furthermore, as the linkages between digital and physical systems (or the 'Internet of Things') expand, the potential for to an adversary to use cyberattacks in *both* kinetic and non-kinetic attacks will increase. A significant risk variable in the operation of autonomous systems is the time that passes between a system failure (i.e. performing in a manner other than how the human operator intended) and the time it takes for a human operator to take corrective action. If the system failure is the result of a deliberate act, this time frame will be compressed (Johnson, 2019a).

Even if nuclear early-warning systems might eventually detect the subversion, heightened levels of uncertainty and tension caused by an alert may impel the respective militaries to put their nuclear weapons on high alert status. This skewed assessment by states in the context of nuclear weapons, which are ready to launch nuclear weapons at a moment's notice, would likely precipitate worst-case scenario thinking that may spark an inadvertent escalation (Talmadge 2017).[28] During a crisis, the inability of a state to determine an attacker's intent may lead an actor to conclude that an attack (threatened or actual) was intended to undermine its nuclear deterrent.[29] For example, an AI-enabled third party generated 'deepfake' (audio and video manipulation), coupled with

[27] AI machine learning systems rely on high-quality data sets to train their algorithms, thus injecting so-called 'poisoned' data into those training sets which could lead these systems to perform in undesired, and potentially undetectable ways.

[28] China's 'no-first-use' nuclear policy, for example, has been a long-standing source of ambiguity and debate. See Talmadge, "Would China Go Nuclear? Assessing the Risk of Chinese Nuclear Escalation in a Conventional War with the United States" (2017).

[29] For example, even if the malware detected in an attack was only capable of espionage, a target may fear that it also contained a 'kill switch' able to disable an early-warning system after activation.

data-poisoning cyberattacks, could spark an escalatory crisis between two (or more) nuclear states (Fitzpatrick 2019).[30]

How could AI-augmented cyber capabilities create new pathways for accidental or inadvertent escalation? To illustrate these dynamics: State A launches a malicious AI-enhanced cyberattack to spoof State B's AI-enabled autonomous sensor platforms and automated target recognition systems, in such a way that the weapon system (e.g. a human-supervised ATR system) is fooled into interpreting a civilian object as a military target. B in response, based on manipulated or erroneous information, and the inability of human supervisors to detect the spoofed imagery that fooled the weapons' automated target recognition algorithm in time to take corrective action, accidentally (and unintentionally) escalates a situation (Libicki 2018). In this example, the spoofing attack on the weapon systems' algorithm is executed in such a way that the imagery appears to the recognition system as indistinguishable from a valid military target (Rohrbach, Hendricks, Burns, Darrell and Saenko 2018) [31] escalating a situation based on a false premise that would unlikely to fool the human eye[32].

Also, the explainability (or 'black box') problem associated with AI applications could compound these potential escalation dynamics.[33] Insufficient understanding of how and why AI algorithms reach a particular judgment or decision might complicate the task of determining whether data sets had been deliberately compromised to manufacture false outcomes (e.g. attacking incorrect targets or misdirecting allies during combat).[34]Moreover, as humans and AI team up to accomplish particular missions, the opacity associated with how AI systems reach a decision may cause an

[30] A recent workshop hosted by IISS demonstrated that potentially malign manipulation of input data received by early-warning systems might not only subvert the output of AI systems in a specific situation, but also undermine the reliability of an entire algorithm network environment if executed during the program's training phase (Fitzpatrick 2019).

[31] AI experts have proven that even when data appears accurate to AI image recognition software, these systems often 'hallucinate' objects that do not exist. See, Rohrbach, Anna., Hendricks, Lisa Anne., Burns, Kaylee., Darrell, Trevor., and Saenko, Kate. "Object Hallucination in Image Captioning" (2018).

[32] This fictional illustration demonstrates one possible (and worst case) outcome of a malicious operation against AI-enhanced weapons systems in an adversarial and human-machine collaborative scenario. Alternative outcomes could be conceived whereby advances in detection algorithms and sensor technology enable a human operator to avert a crisis before military force is used. Also, other safety mechanisms (i.e. circuit breakers or redundancies built into the systems) might also prevent this situation from spiralling.

[33] 'Explainability' is a term used by AI experts referring to the fact that many AI systems produce outcomes with no explanation of the path the system took to derive the solution (Johnson, 2019a).

[34] For example, it is uncertain how independently integrated AI platforms might interact with one another in a cross-domain battlefield environment.

operator to have either too much or too little confidence in a system's performance. As a result, unless the system's machine learning algorithm is terminated, at the end of the training phase, once deployed, it could potentially learn something it was not intended to, or even perform a task or mission that its human designers do not expect it to do.[35]

In sum, technologies that improve the reliability and speed with which information is processed and disseminated by early-warning systems may paradoxically also increase the vulnerabilities of these networks, creating new first-mover advantages and escalation pathways during a crisis that spark unintentional (or intentional) warfare (Schneider 2019).[36] Put another way, AI might be developed in order to enable a state to pursue a pre-determined escalatory path. Thus, increased escalation risk as a result of technology is *not always* inadvertent or accidental (Long and Rittenhouse Green 2014).[37]

'Cyber guns' supercharged with AI machine learning

While manipulations and system subversions are possible with existing cyber offence tools, rapid advances in AI and increasing degrees of military autonomy could amplify the potential speed, power and scale of future attacks in cyberspace.[38] There are three significant intersections between AI systems and cybersecurity that are most salient to the military domain: (1) advances in autonomy and machine learning techniques mean that a much broader range of physical systems are now vulnerable to cyberattacks (i.e. hacking and data-poisoning) (Greenberg 2016)[39]; (2) cyberattacks on AI systems can offer attackers access to machine learning algorithms, trained models being used by the application and a potential vast amounts of data from facial recognition and intelligence collection and analysis systems (e.g. satellite navigation and imagery systems used to cue precision munitions and supported intelligence, surveillance, reconnaissance

[35] This problem set is one of the main reasons why the use of AI machine learning in the context of weapon systems has, thus far, been limited to experimental research. See, Boulanin and Verbruggen. "Mapping the Development of Autonomy in Weapon Systems" (2017).

[36] This paradox suggests that when states' capabilities are dependent resources (e.g. manpower or data sets) that can be exploited or controlled by an adversary, both sides have incentives for first strike (Schneider 2019).

[37] For example, heightened escalation risks caused by aggressive US-Soviet expansion of counterforce technology during the Cold War reflected shifting nuclear doctrines on both sides (i.e. away from assured mutual destruction), not the pursuit of these technologies themselves (Long and Rittenhouse Green 2014). "

[38] This vulnerability is particularly concerning given the dual-use nature of AI technology.

[39] For example, a hacker brought a Jeep to a standstill on a busy highway and was then able to interfere with its steering system to cause unintended acceleration (Greenberg 2016).

missions); and (3) the use of AI systems in conjunction with existing cyber offence tools will enable sophisticated cyberattacks to be executed at larger scale (both geographically and across networks), at faster speeds, and simultaneously across multiple military domains with improved anonymity. In short, despite the relatively benign augmentation mechanisms used to develop cyber offensive capabilities with AI systems, the resultant impact on the speed and scope of AI cyber tools will likely have destabilizing effects, nonetheless[40].

Recent advances in machine learning have significantly contributed to resolving several technical bottlenecks in several fields of AI, which could allow for significant qualitative improvements to a large variety of autonomous weapon systems and applications. In combination, machine learning and autonomy could intersect with nuclear security in a multitude of ways, with *both* positive and negative implications for strategic stability. For example, machine learning could enhance the detection capabilities of (nuclear and non-nuclear) early-warning systems and improve the collection and cross-analysis of ISR information, thus reducing the risk of inadvertent and accidental escalation. Machine learning may also bolster the cyber-security of command and control systems, create new and possibilities for arms control, testing, verification and monitoring nuclear stockpiles, and enhance conventional counterforce capabilities. Unmanned autonomous systems could be deployed in complex missions in hitherto inaccessible and cluttered environments (e.g. under-sea anti-submarine warfare) and aerial and underwater drones in swarms might eventually replace intercontinental ballistic missiles (ICBMs) and nuclear-armed submarines (SSBNs) for the delivery of nuclear weapons.

In response to these anticipated vulnerabilities, and to achieve the (perceived) first-mover advantages the increased speed in AI-augmented cyber tools could confer, China, Russia and the United States have continued to harden their AI cyber defences. Despite these parallel efforts, divergent views exist on how strategic communities perceive the nature of the risks posed by AI-enhanced cyberattacks against dual-use C3I systems. Thus, AI-augmented cyber intelligence-gathering tools targeting an adversary's command and control assets (especially dual-use facilities) during a crisis may be misinterpreted as a prelude to an imminent pre-emptive attack on its nuclear

[40] Many AI additions that can augment offensive cyber capabilities involve either enumerating the target space or repackaging malware to avoid detection.

force and therefore as undermining strategic stability.

Moreover, AI machine learning could enable new offensive concepts such as a so-called 'left-of-launch operation' (Riki 2015), [41] thereby compounding the ambiguities and fear about pre-emptive attacks in cyberspace, and increasing first-mover incentives – or use-them-or-lose-them situations (Buchanan 2017; Buchanan and Miller 2017)[42]. In the cyber domain, it is generally unclear in the early stages of a cyber operation whether an adversary intends to collect intelligence or prepare for an offensive attack, which is more likely to be used early on in a conflict scenario. Open sources suggest, for example, that Chinese analysts view the vulnerability of China's nuclear command, control, and communications (C3I) systems to cyber infiltrations – even if an attacker's objective was limited to cyberespionage – as a highly escalatory national security threat. By contrast, Russian analysts tend to view Russia's nuclear C3I network as relatively isolated, and thus insulated from cyberattacks.

Irrespective of these differences, the uncertainty caused by the possibility, actual or otherwise, that AI-enhanced offensive cyber tools might be used (or threatened) to undermine the credibility and reliability of states' nuclear forces would be destabilizing, both as result of the existence of enhanced cyber offensive capabilities and the potential efficacy of these tools against nuclear forces. As the historical record attests, rational actors have incentives to misrepresent what they know (i.e. lie and bluff) in order to improve the terms of any settlement that may emerge from the wartime bargaining process – or the risk-reward payoff. Even a modicum of uncertainty about the effectiveness of AI-augmented cyber capabilities during a crisis or conflict would, therefore, reduce *both* sides' risk tolerance and increase the incentive to strike pre-emptively as a hedging strategy.

A virtuous cycle that might flow from enhanced reassurances premised on comprehensive intelligence would require information symmetry (i.e. equal access to intelligence and analysis systems) between great and rising powers, and shared confidence in the accuracy and credibility of these systems. Perhaps most challenging of all in a world of 'revisionist' rising powers, the intentions of all rival states would need to be genuinely benign for this optimistic view to occur. Under crisis conditions, for example, an offensive AI cyber tool that succeeds in compromising an adversary's

[41] A 'left-of-launch operation' refers to a cyber-offensive operation that would defeat the threat of a nuclear ballistic missile before it is launched (Riki 2015).
[42] For more on this topic see Buchanan (2017) and Buchanan and Miller (2017).

nuclear weapon systems, resulting in an 'asymmetric information' situation, could cause either or both sides to overstate (or understate) its retaliatory capabilities, and in turn, be more inclined to act in a risky and escalatory fashion. In short, in a competitive strategic environment, where states are inclined to assume the worst of others' intentions, one state's efforts to enhance the survivability of its strategic forces may be viewed by others as a threat to their nuclear retaliatory capability – or second-strike capacity.

Manipulation of the information landscape

While machine learning big-data analytics, sensing technology, networks supported by 5G networks, could alert commanders of incoming threats with increased speed and precision, and make accidents caused by human error potentially less likely (especially accidents and malfunctions caused by false warning), these advances could simultaneously amplify escalation risks in two ways.[43]

First, AI machine learning used as force multipliers for cyber offence (e.g. data poisoning spoofing, 'deepfakes,' manipulation, hacking and digital jamming) would be considerably more difficult to detect – especially if an attacker used AI advanced persistent threat (APT) tools – or 'hunting for weaknesses' in spectrum-contested environments. Second, in the unlikely event an attack was successfully detected, threat identification (or attribution) at machine speed would be virtually impossible. In addition to the co-mingling and the speed of warfare issues, AI machine learning systems might also exacerbate the existing – well-established and non-kinetic – risk of inadvertent (and intentional) escalation from the manipulation of the information landscape in which decisions about nuclear weapons are located. For example, the use of deepfake videos built-in real-time and disseminated online for malicious purposes.

Human cognition, and thus effective deterrence and signalling intentions, is predicated on reliable and clear information; if an adversary is concerned that the information available to them is limited (or worse, inaccurate), they will likely assume the worst and act accordingly. Asymmetric information situations between rivals about the balance of military power could undermine crisis stability, and in turn, create rational incentives to escalate to nuclear confrontation. Consequently, states will be

[43] For example, in 1983, a malfunctioning Soviet early-warning system led to the 'detection' of a non-existent US attack.

more inclined to assume the worst of others' intentions, especially in situations where the legitimacy of the status quo is contested (i.e. maritime Asia).

In the competitive strategic environment, it is easy to imagine unprovoked escalation caused by a malicious third-party (or state-proxy) clandestine action. Even if nuclear early-warning systems might eventually detect the subversion, heightened levels of uncertainty and tension caused by an alert may impel the respective militaries to put their nuclear weapons on high alert status. During a crisis, the inability of a state to determine an attacker's intent may lead an actor to conclude that an attack (threatened or actual) was intended to undermine its nuclear deterrent.[44] For example, in an effort to incite conflict between two rival states, State A uses proxy hackers to use a AI machine learning technique (i.e. generative adversarial networks GAN) to launch 'deepfake' video or audio material, depicting senior military commanders of State B conspiring to launch a pre-emptive strike on State C.[45] Then, this 'deepfake' footage is deliberately leaked into C's AI-augmented intelligence collection and analysis systems, provoking C to escalate the situation with strategic consequences. B responds in kind.[46]

The current underdeveloped state of 'counter-AI' capabilities and other fail-safe mechanisms (e.g. circuit breakers) to de-escalate escalation in cyberspace, will make the unprovoked and unintentional escalation dynamics depicted in this scenario very challenging to mitigate. Moreover, in the emerging 'deepfakes' arms race (much like cybersecurity more generally), detection software will likely lag behind advances in offensive enabling solutions – or offence-dominant ones. According to computer science expert Hany Farid, there are probably 100 to 1000 times 'more people developing the technology to manipulate content than there is to detect [it]' (Fontaine and Frederick 2019).

The overreliance on automation (or 'automation bias') in the deployment of increasing complex AI-augmented capabilities – such as cyber, C3I systems, AWS and

[44] For example, even if the malware detected in an attack was only capable of espionage, a target may fear that it also contained a 'kill switch' able to disable an early-warning system after activation.

[45] The generative adversarial network (GAN) is a new approach that involves two artificial neural network systems that spar with each other to create realistic original image, audio or video content, something machines have never been able to do properly before (Boulanin, 2019).

[46] Similar to the previous fictional case study, alternatives to this worst scenario outcome can also be conceived. At the most basic level, crisis could be averted if the deepfake that led to the catastrophe was detected early enough. The case study also assumes the victims of this offensive operation viewed each other as adversaries, making them more suspicious of the others' intentions in the event of a crisis. Were *either* side able to demonstrate empathy or display restraint, however, the escalatory outcome that followed might have been avoided.

precision missile munitions – designed to provide a distinct tactical advantage of machine-speed, will increase the vulnerability of these capabilities to exploitation. Moreover, the increasing substitution of human cognition for logic derived from machines will likely increase the opportunities for adversaries to exploit the limitations of narrow AI technology (i.e. lack of human intuition, brittleness in complex real-world situations and an inability to effectively detect or counter manipulation attacks (Libicki 2018).[47] Therefore, until researchers unravel some of the unexplainable features of AI, human error and machine error will likely compound one another, with unpredictable results.[48] Simply put, we are at a critical crossroads in the parallel (and symbiotic) evolution of AI and cyberspace that national security communities, globally, will need to prepare for proactively.

Policy interventions

A prominent theme that runs through this article – and central to understanding the potential impact AI for strategic stability and nuclear security more broadly – is the concern that AI systems operating at machine-speed will push the pace of combat to a point where the actions of machines surpass the (cognitive and physical) ability of human decision makers to control (or even comprehend) events.

Possible multi-track policy responses to push back against the threat posed to stability from AI in a multipolar can be broadly categorized into: (1) those that focus on enhancing debate and discussion between researchers, global defence communities, decision makers, academics and other political and societal stakeholders; and (2) a range of specific policy recommendations and tools for great military powers to negotiate and implement.

Enhancing debate and dialogue

First, to mitigate (or at least manage) the destabilizing and escalatory risks posed by the AI-cybersecurity nexus, great military powers must closely coordinate their confidence-building measures in an effort to pre-empt some of the risks to stability outlined above. Track-1 and track-2 discussions should include diplomats and military

[47] One potential countermeasure could be to train each AI system in a slightly different way, to limit the success of system failures caused by exploitation to result in the collapse of an entire overall military operation. See Libicki (2018).

[48] Generally, the less predictable the environment, the harder it is to model, and thus, the more difficult it becomes to create autonomous capabilities within systems that are effective, safe and reliable.

leaders (especially from China, Russian and the United States), industry experts, AI researchers and multi-disciplinary academics. Specifically, great military powers should establish an international framework for governance, norms, behaviour and the regulation of policy on machine learning-augmented cyber capabilities. Resistance to these efforts will likely come from states who worry that in revealing their cyber (especially offensive) capabilities, they could upend the deterrence utility of these tools.

To be sure, the challenges posed in the coordination and implementation of these policies will require bold and visionary leadership to circumvent the inevitable regional agendas, interdisciplinary resistance and burgeoning security dilemmas between rival states. Because of the rapid technological change in AI formal treaties, associated with arms-control agreements that require lengthy and complex negotiation and ratification processes, legal frameworks risk becoming obsolete before they come into effect. The historical record has demonstrated on several occasions that these kinds of challenges facing humanity can be overcome (Naur and Randell 1968; Krimsky 1962).[49]

Second, the think tank community, academics and AI research experts should pool their resources to investigate the implications of the AI-cyber nexus for a range of potential security issues such as: the impact of AI bias on future military-use applications; how to prepare for (and react to) artificial general intelligence; and measures to mitigate, prevent and manage offensive uses of AI.[50] If necessary, steps could be taken to amend existing legal definitions of offensive cyber operations, such as hacking, data-poisoning attacks and spoofing, to account for the increased proliferation and potential damage caused by AI.

Third, and related, due to the intrinsic dual-use nature of the AI-cyber security problem set, this dialogue should also be expanded to include other stakeholders such as private-sector AI and cybersecurity experts, the commercial sector, ethicists, philosophers, civil society and public opinion.[51] States should also collaborate on dual-use AI research to leverage AI's low-cost and scaling advantages (i.e. in autonomy and

[49] Examples include: the 1968 NATO conference at Garmisch, which established a consensus on the mounting risks from software systems and the 1975 NIH conference at Asilomar that underscored the risks posed by recombinant DNA research. See Naur and Randell (1962).

[50] AI-related research findings are often not made public due to reasons related to intellectual property and broader national security concerns.

[51] It is important to recognize that different communities and cultures will have varying abilities to make such adaptations (e.g. technological literacy, culture norms and economic systems) which may pose challenges for implementing security policies in society at large.

robotics). Further, a focus on the safety, testing and robustness of AI systems is a critical step in mitigating potential vulnerabilities and risks caused by errors, bias, and explainability in uncontrolled and complex environments.[52] In the case of offensive AI-related cybersecurity research, where public distribution might cause vulnerabilities and worsen security, publication could be confined to trusted organizations and entities.

Best practices that exist in more mature methods for addressing dual-use concerns, like computer security, where applicable, might be applied to AI. For example, the wide use of red teaming exercises to enhance network security, organization and practices. Specifically, AI-cyber red teaming – like DARPA's Cyber Grand Challenge – will enable engineers and operators to better understand the skills needed to execute particular offensive and defensive operations (especially using machine learning techniques), and to better manage system vulnerabilities, adversarial exploitation, stress-testing and social engineering challenges (Anderson et al. 2016).

A recent study explored case studies of previous examples (biological weapons, cryptography and nuclear technology) of dual-use technologies – with civilian and military uses – to examine potential insights for AI dual-use risk management policies such as export controls and pre-publication reviews. These insights also highlight the potential challenges of establishing regulatory, legal and normative frameworks for dual-use technologies – for example, the cautionary tale of ineffective efforts, in the late 1990s, to regulate cryptographic algorithms and cyber-network security tools through export controls.

The extent to which the integration of AI into dual-use weapon systems (such as cybersecurity) might influence actors' attitude to risk, the offence-defence balance and, in turn, perceptions of others' intentions and capabilities, could have profound implications for strategic deterrence and nuclear security.[53] In sum, cognizant that some states have deployed (or imminently plan to deploy) AI systems, experts generally agree that AI requires further experimentation, testing and development before being integrated into lethal weapon systems and their decision-making support systems.

[52] Despite the improvements made to reinforcement learning techniques to handle increasingly complex and uncertain real-world situations, currently, algorithms function as a result of a learning process derived from outcome data defined by humans. Thus, AI systems remain of limited use in complex combat contexts where the outcome unclear.

[53] The White House's 2016 'Partnership on AI,' series of workshops on AI and the 2017 'Beneficial AI' conference in Asilomar, and the 'AI Now' conference series are good examples of this kind of research collaboration.

Taking the lead from the cybersecurity community

Some examples of cybersecurity-centered measures that might be further researched and implemented to pre-empt and mitigate some of risks posed by AI-augmented cyberattacks (especially involving safety-critical nuclear systems) highlighted in this paper include the following.

First, coordinating AI-simulated war games, red teaming creative thinking exercises and creating redundancies (i.e. back-ups or fail-safes) in networks to detect errors, fix vulnerabilities and increase the reliability and robustness of military (especially nuclear-centric) systems. Specific procedures could be implemented to enable confidential reporting and fixing the vulnerabilities, subversions and other kinds of manipulations detected in AI systems. These findings could be used to track the proliferation of AI-related incursions and then countermeasures could be developed and standardized to manage these threats[54].

Second, states should formalize verification methods and protocols and consider issues such as: to what extent, under what circumstances, and for what types of AI systems can formal verification be implemented? Might other approaches be developed to achieve similar goals (e.g. machine learning and big data analysis augmented verification methods)? For example, DARPA's Assured Autonomy Program combines data-driven machine learning algorithms to assure the safety of autonomous cyber physical systems (air, ground, sea and undersea unmanned vehicles), which continue to learn throughout their lifespans.

This learning characteristic makes assurance or verification using traditional methods especially challenging. Whether AI applications in the military domain can be formally verified, for now, is an unanswered question.[55] Much like that of cyber systems, the complexity of AI systems and the difficulty of defining their properties for formal verification makes them less amenable to verification compared to other types of technology.

Third, the global defence communities should actively invest in the development of AI cyber-defence tools (e.g. analysing classification errors, automatic

[54] The technical and practical feasibility of any AI-related security-enhancing hardware or software would also need to be considered by programmers and users.

[55] Given that broader factors and stakes will likely be involved as a situation moves up the rungs of the escalation ladder (especially political and strategic/doctrinal contexts), these kinds of technical restraints and controls may not necessary reduce the risk of conflict.

detection of remote vulnerability scanning and model extraction improvements), AI-centric secure hardware, and other fail-safe mechanisms, to allow for de-escalation and to prevent unintentional or accidental escalation.

Several issues that need further exploration related to these measures include:[56] the extent to which existing tools might be effective against vulnerabilities in AI systems? How can these tools be tailored for AI systems across multiple military domains? Is there an equivalent to 'patching' in military AI systems? What kinds of policies might incentivize, and ensure compliance with, meaningful reforms to existing hardware in the military sphere? While these questions are challenging, and necessarily speculative for now, answers will become more evident as the technology matures.

Disclosure statement: No potential conflict of interest was reported by the author.

Bibliography

Acton, James M., Li Bin., Alexey Arbatov., Petr Topychkanov., and Zhao Tong. 2017. *Entanglement: Russian and Chinese Perspectives on Non-Nuclear Weapons and Nuclear Risks*. Washington DC: Carnegie Endowment for International Peace.

Allen, Greg and Chan, Taniel. 2017. *Artificial Intelligence and National Security*. Cyber Security Project: Belfer Center. Cambridge, MA: Harvard Kennedy School.

Anderson, Hyrum., Jonathan Woodbridge and Bobby Filar. 2016. "DeepDGA: Adversarially-Tuned Domain Generation and Detection." *Arxiv.* https://arxiv.org/abs/1610.01969

Ayoub, Kareem and Payne, Kenneth. 2016. "Strategy in the Age of Artificial Intelligence," Journal of Strategic Studies 39 (5-6): 799-805.

Berlinger, Joshua and Juliet Perr. 2017. "China Tried to Hack Group Linked to Controversial Missile Defense System, U.S. Cybersecurity Firm Says." *CNN.* Accessed May 5, 2018. https://www.cnn.com/2017/04/27/asia/china-south-korea-thaad-hack/index.html

Biddle, Stephen D., and Robert Zirkle. 1996. "Technology, Civil-Military Relations, and Warfare in the Developing World." *Journal of Strategic Studies* 19 (2): 171-212.

Boden, Margaret. A. 2016. *AI Its Nature and Future*. Oxford: Oxford University Press.

Boulanin, Vincent, ed. 2019. *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Vol. I Euro-Atlantic Perspectives.* Stockholm: SIPRI Publications.

Boulanin, Vincent and Verbruggen, Maaike, ed. 2017. *Mapping the Development of Autonomy in Weapon Systems.* Stockholm: SIPRI Publications.

Brodie, Bernard. 1959. *Strategy in the Missile Age*. Princeton, NJ: Princeton University Press.

Brodie, Bernard, and Fawn Brodie. 1973. *From Crossbow to H-Bomb*. Bloomington: Indiana. University Press. Chapter 23.

Buchanan, Ben. 2017. *The Cybersecurity Dilemma*. New York: Oxford University Press.

Buchanan, Ben and Taylor Miller. 2017. "Machine Learning for Policymakers." Paper, Cyber Security Project: Belfer Center. Cambridge, MA: Harvard Kennedy School.

Conklin, Ben. 2018. "How artificial intelligence is transforming GEOINT." *GCN*. Accessed April 20, 2018. https://gcn.com/articles/2018/04/18/ai-transform-geoint.aspx

Connable, Ben. 2012. *Embracing the Fog of War: Assessment and Metrics in Counterinsurgency*. Santa Monica, CA: RAND Corporation.

Dvorsky, George. 2017. "Hackers Have Already Started to Weaponize Artificial Intelligence." *Gizmodo*.
Accessed May 6, 2018. https://gizmodo.com/hackers-have-already-started-to-weaponize-artificial-in-1797688425

Ellison, Riki. 2015. "Left of Launch." *Missile Defense Advocacy Alliance*. Accessed June 6, 2019. https://missiledefenseadvocacy.org/alert/3132/

Gartzke, Erik, and Jon R. Lindsay. 2015. "Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace." *Security Studies* 24 (2): 316-348.

Gartzke, Erik, and Jon R. Lindsay. 2017. "Thermonuclear Cyberwar." *Journal of Cybersecurity* 3 (1): 37-48.

Gartzke, Erik, and Jon R. Lindsay, eds. 2019. *Cross-domain Deterrence: Strategy in an Era of Complexity*. Oxford: Oxford University Press.

Fearon, James D. 1995. "Rationalist Explanations for War." *International Organization* 49 (3): 379-414.

Fisher, Kathleen. 2014. “Using Formal Methods to Enable More Secure Vehicles: DARPA’s HACMS Program.” ICFP 2014: Proceedings of the 19th ACM SIGPLAN International Conference on Functional Programming.

Fitzpatrick, Mark. 2019. “Artificial Intelligence and Nuclear Command and Control.” *Survival* 61 (3): 81-92.

Fontaine, Richard, and Kara Frederick. 2019. “The Autocrat’s New Tool Kit.” *The Wall Street Journal.* Accessed March 16, 2019. https://www.wsj.com/articles/the-autocrats-new-tool-kit-11552662637

Gasser, Paige., Rafeal Loss and Andrew Reddie. 2018.*Assessing the Strategic Effects of Artificial Intelligence – Workshop Summary.* California: Center for Global Security Research Lawrence Livermore National Laboratory.

Glaser, Charles, and Chaim Kaufmann. 1998. “What Is the Offense-Defense Balance and How Can We Measure It?” *International Security* 22 (4): 44-82.

Gray, E. et al. 2015. “Small Big Data: Using Multiple Data Sets to Explore Unfolding Social and Economic Change.” *Big Data & Society* 2 (1): 1-6.

Greenberg, Andy. 2016. “The Jeep Hackers are Back to Prove Car Hacking Can Get Much Worse.” *Wired*. Accessed May 10, 2019. https://www.wired.com/2016/08/jeep-hackers-return-high-speed-steering-acceleration-hacks/

Hao, Karen. 2019. “Inside the World of AI that Forges Beautiful Art and Terrifying Deepfakes.” *MIT Technology Review.*
Accessed January 10, 2019. https://www.technologyreview.com/s/612501/inside-the-world-of-ai-that-forges-beautiful-art-and-terrifying-deepfakes/

Hoadley, Daniel S., and Lucas J. Nathan. 2018. *Artificial Intelligence and National Security*. Washington DC: Library of Congress, Congressional Research Service.

Hoffman, Wyatt. 2019. “Is Cyber Strategy Possible?” *The Washington Quarterly* 42 (1): 131-152.

Horowitz, Michael C. 2018. “Artificial Intelligence, International Competition, And The Balance of Power.” *Texas National Security Review* 1 (3): 37-57.

Ingram, Paul. 2017. “Hacking UK Trident: A Growing Threat.” *BASIC.*
Accessed September 10, 2017. https://basicint.org/publications/stanislav-abaimov-paul-ingram-executive-director/2017/hacking-uk-trident-growing-threat

Jervis, Robert. 2016. “Some Thoughts on Deterrence In The Cyber Era.” *Journal of Information Warfare* 15 (2): 66-73.

Johnson, James. 2019a “Artificial Intelligence & Future Warfare: Implications For International Security.” *Defense & Security Analysis* 35 (2): 147-169.

Johnson, James. 2019b "The End of Military-Techno Pax Americana? Washington's Strategic Responses to Chinese AI-enabled Military Technology." *The Pacific Review* DOI: 10.1080/09512748.2019.1676299

Krimsky, Sheldon. 1962. *Genetic Alchemy: The Social History of the Recombinant DNA Controversy*. Cambridge, MA: MIT Press.

Kristensen, H. M., M. McKinzie, and T. A. Postol. 2017. "How US nuclear force modernization is undermining strategic stability: the burst-height compensating super-fuze." *Bulletin of the Atomic Scientists*.
Accessed June 5, 2018. https://thebulletin.org/2017/03/how-us-nuclear-force-modernization-is-undermining-strategic-stability-the-burst-height-compensating-super-fuze/

Lee, Connie. 2019. "SOCOM Plans New Artificial Intelligence Strategy." *National Defense*.
Accessed September 5, 2019.
https://www.nationaldefensemagazine.org/articles/2019/8/9/socom-plans-new-artificial-intelligence-strategy

Lewis, Jeffrey. 2016. "Is Launch Under Attack Feasible?" *Nuclear Threat Initiative*. Accessed September 6, 2018 https://www.nti.org/analysis/articles/launch-under-attack-feasible/

Lewis, Patricia, and Beyza Unal. 2019. *Cybersecurity of nuclear weapons systems: Threats, vulnerabilities and consequences*. London, United Kingdom: The Royal Institute of International Affairs, Chatham House.

Lieber, Keir A., and Daryl G. Press. 2013. "Why States Won't Give Nuclear Weapons to Terrorists." *International Security*. 35 (8): 80-104.

Libicki, Martin. 2016. *Cyberspace in Peace and War*. Annapolis: Naval Institute Press.

Libicki, Martin. 2018. "A Hacker Way of Warfare." Chapter 18 in *AI, China, Russia, and the Global Order: Technological, Political, Global, and Creative Perspective*, 128-132. Wright D. Nicholas, ed.. Washington, DC: A Strategic Multilayer Assessment Periodic Publication.

Lin, Herbert. 2016. "Reflections on the New Department of Defense Cyber Strategy: What It Says, What It Doesn't Say." *Georgetown Journal of International Affairs* 17 (3): 5-13.

Lin, Herbert, and Amy Zegart, eds. 2019. *Bombs, Bytes and Spies: The Strategic Dimensions of Offensive Cyber Operations*. Washington, DC: Brookings Institution.

Lobo, Savia. 2019. "The U.S. DoD wants to dominate Russia and China in Artificial Intelligence. Last week gave us a glimpse into that vision." *Packt*.

Accessed April 4, 2019. https://hub.packtpub.com/the-u-s-dod-wants-to-dominate-russia-and-china-in-artificial-intelligence-last-week-gave-us-a-glimpse-into-that-vision/

Long, Austin, and Brendan Rittenhouse Green. 2014. “Stalking the Secure Second Strike: Intelligence, Counterforce, and Nuclear Strategy.” *Journal of Strategic Studies* 38 (1-2): 38-73.

Lynn-Jones, Sean. 1995. “Offense-Defense Theory and Its Critics.” *Security Studies* 4 (4): 660-691.

Maurer, Tim. 2017. *Cyber Mercenaries: The State, Hackers, and Power*. New York, NY: Cambridge University Press.

Metz, Steven. 2000. *Armed Conflict in the 21st Century: The Information Revolution and Post-Modern Warfare*. Carlisle: Strategic Studies Institute.

Metz, Steven, and James Kievit. 1995. *Strategy and the Revolution in Military Affairs: From Theory to Policy*. Carlisle: Strategic Studies Institute.

Moore, Andrew W. 2017. “AI and National Security in 2017.” Paper presented at the AI and Global Security Summit, Washington, DC, November 1.

Naur, Peter, and Brian Randell. 1968. *Software Engineering: Report on a Conference Sponsored by the NATO Science Committee.* Brussels: Scientific Affairs Division, NATO.

Nicholas D. Wright, ed. *A Strategic Multilayer Assessment (SMA) Periodic Publication*. (Washington, DC: SMA Publications).

Nilsson, Nils J. 2010. *The Quest for Artificial Intelligence*. (New York: Cambridge University Press).

Osborn, K. 2018. “Navy Cyber War Breakthrough – AI Finds Malware in Encrypted Traffic.” *Warrior Maven.*
Accessed January 5, 2019. https://defensemaven.io/warriormaven/cyber/navy-cyber-war-breakthrough-ai-finds-malware-in-encrypted-traffic-K_tLobkkJkqadxDT9wPtaw/

Pearl, Andy. 2017. “Homage to John McCarthy, the father of artificial intelligence (AI).” *Artificial Solutions*.
Accessed January 8, 2019. https://www.artificial-solutions.com/blog/homage-to-john-mccarthy-the-father-of-artificial-intelligence

Preparing for the Future of Artificial Intelligence. 2016. National Science and Technology Council, Committee on Technology. October 12. Washington, DC: Executive Office of the President.

Reedy, Christianna. 2017. “Get ready for humanity 2.0.” *Future Society*.
Accessed January 26, 2019. https://futurism.com/kurzweil-claims-that-the-singularity-

will-happen-by-2045

Rid, Thomas. 2012. “Think Again: Cyberwar.” *Foreign Policy.* Accessed January 20, 2019. https://foreignpolicy.com/2012/02/27/think-again-cyberwar/

Rid, Thomas. 2013. *Cyber War Will Not Take Place*. New York: Oxford University Press.

Rohrbach, Anna, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. “Object Hallucination in Image Captioning.” *Cornell University Library*.

Russel, S., and P. Norvig. 2014. *Artificial Intelligence: A Modern Approach*. 3rd ed. Harlow: Pearson Education.

Schelling, Thomas C., and Morton Halperin. 1961. *Strategy and Arms Control*. New York, NY: The Twentieth Century Fund.

Schmidhuber, Jurgen. 2015. “Deep Learning in Neural Networks: An Overview.” *Neural Networks* 61: 85-117.

Schneider, Jacquelyn. 2019. “The Capability/Vulnerability Paradox and Military Revolutions: Implications for Computing, Cyber, and the Onset of War.” *Journal of Strategic Studies* 42 (6):841-863.

Sechser, Todd S., Neil Narang, and Caitlin Talmadge. 2019. “Emerging Technologies and Strategic Stability in Peacetime, Crisis, and War.” *Journal of Strategic Studies*, 42 (6): 727-735.

Shachtman, Noah. 2011. “Exclusive: Computer virus hits U.S. drone fleet.” *Wired*. Accessed June 15, 2019 https://www.wired.com/2011/10/virus-hits-drone-fleet/

Shehadeh, Karim K. 1999. “The Wassenaar Arrangement and Encryption Exports: An Ineffective Export Control Regime that Compromises United States Economic Interests.” *American University of International Law Review* 15 (1): 271-319.

Slayton, Rebecca. 2017. “What is Cyber Offense-Defense Balance?” *International Security* 41 (3): 72-109.

Talmadge, Caitlin. 2019. “Emerging Technology and Intra-war Escalation Risks: Evidence From The Cold War, Implications for Today.” *Journal of Strategic Studies* 42 (6): 864-887.

Talmadge, Caitlin. 2017. “Would China Go Nuclear? Assessing the Risk Of Chinese Nuclear Escalation In A Conventional War With The United States.” *International Security* 41 (4): 50-92.

Neema, Sandeep. 2017. “Assured Autonomy.” *DARPA*. Accessed April 7, 2018. https://www.darpa.mil/program/assured-autonomy

The White House. 2017. *National Security Strategy of the United States of America*.

University of Oxford, University of Cambridge, Future of Humanity Institute, OpenAI & Future of Humanity Institute. 2018. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Oxford: Oxford University Press.

US Department of Defense. 2019. “Summary of the 2018 Department of Defense Artificial Intelligence Strategy.”
Accessed June 1, 2019. https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

Vernon, David. 2014. *Artificial Cognitive Systems: A Primer*. Cambridge, MA: MIT Press.

Wright, Peter. 2019. “AI, China, Russia, and the Global Order: Technological, Political, Global, and Creative Perspectives.” *NSI*.
Accessed June 10, 2019. https://nsiteam.com/social/wp-content/uploads/2018/12/AI-China-Russia-Global-WP_FINAL.pdf

Zarkadakis, George. 2015. *In Our Image*. New York: NY: Pegasus Books.