2020 12th International Conference on Cyber Conflict
20/20 Vision: The Next Decade
T. Jančárková, L. Lindström,
M. Signoretti, I. Tolga, G. Visky (Eds.)
2020 © NATO CCDCOE Publications, Tallinn

Problems of Poison: New Paradigms and "Agreed" Competition in the Era of AI-Enabled Cyber Operations

Christopher Whyte
L. Douglas Wilder School of Government and Public Affairs
Virginia Commonwealth University

Abstract: Few developments seem as poised to alter the characteristics of security in the digital age as the advent of artificial intelligence (AI) technologies. For national defense establishments, the emergence of AI techniques is particularly worrisome, not least because prototype applications already exist. Cyber attacks augmented by AI portend the tailored manipulation of human vectors within the attack surface of important societal systems at great scale, as well as opportunities for calamity resulting from the secondment of technical skill from the hacker to the algorithm. Arguably most important, however, is the fact that AI-enabled cyber campaigns contain great potential for operational obfuscation and strategic misdirection. At the operational level, techniques for piggybacking onto routine activities and for adaptive evasion of security protocols add uncertainty, complicating the defensive mission particularly where adversarial learning tools are employed in offense. Strategically, AI-enabled cyber operations offer distinct attempts to persistently shape the spectrum of cyber contention may be able to pursue conflict outcomes beyond the expected scope of adversary operation. On the other, AI-augmented cyber defenses incorporated into national defense postures are likely to be vulnerable to "poisoning" attacks that predict, manipulate and subvert the functionality of defensive algorithms. This article takes on two primary tasks. First, it considers and categorizes the primary ways in which AI technologies are likely to augment offensive cyber operations, including the shape of cyber activities designed to target AI systems. Then, it frames a discussion of implications for deterrence in cyberspace by referring to the policy of persistent

engagement, agreed competition and forward defense promulgated in 2018 by the United States. Here, it is argued that the centrality of cyberspace to the deployment and operation of soon-to-be-ubiquitous AI systems implies new motivations for operation within the domain, complicating numerous assumptions that underlie current approaches. In particular, AI cyber operations pose unique measurement issues for the policy regime.

Keywords: *deterrence, persistent engagement, cyber, AI, machine learning*

1. INTRODUCTION

In recent decades, few technological developments have captured the attention and sparked the concern of national publics as much as those linked to artificial intelligence (AI). This might seem a remarkable and outlandish statement, given that, if prompted, the median consumer would likely be unable to identify that AI sits at the heart of everyday commercial services like Google's search engine or Amazon's marketplace. Nevertheless, the subject of AI has, since at least 2017, come to sit at the heart of prominent conversations about the future of human innovation and the changing shape of societal security.[1] Tech luminaries continue to expound the revolutionary potential of new machine learning and reasoning techniques which now easily solve those endemic issues of over-complexity that plague the conventional design and operation of digital systems. At the same time, leading voices – from Elon Musk to Max Tegmark and Steve Wozniak – increasingly refuse to disagree with doomsayers who claim that AI might, if mismanaged, lead to societal disaster.[2] Indeed, some are so concerned that they lean heavily into threat inflation, using extreme examples – such as the well-publicized threat of autonomous machine "slaughter bots" that, in a fictional future, catalyze societal breakdown as governments and private actors alike are empowered to kill opponents anonymously and at scale[3] – in an attempt to convince audiences of the stakes involved in getting AI "right".[4]

1 It should be noted that the topic of AI involved in the organization and application of military functions is not new, particularly in popular media. Instances of storytelling and more factual exploration can be found in film and written work stretching back through the early-mid 20th century.

2 See, among others, S. Hawking, S. Russell, M. Tegmark, and F. Wilczek, "Transcendence Looks at the Implications of Artificial Intelligence - But Are We Taking AI Seriously Enough?" *The Independent*, January 5, 2014; and Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017).

3 Jessica Cussins. "AI Researchers Create Video to Call for Autonomous Weapons Ban at UN," *The Future of Life Institute*, accessed 28 November 2017, https://futureoflife.org/2017/11/14/ai-researchers-create-video-call-autonomous-weapons-ban-un/.

4 For an overview of expert opinion on AI, see Vincent C. Müller and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in *Fundamental Issues of Artificial Intelligence*, ed. Vincent C. Müller (Synthese Library; Berlin: Springer, 2016), 555-72.

Around the world, few entities are as focused on the impact that AI systems portend for security as national militaries. In the United States, political and military leaders have variously called for a "Third Offset" that leverages smart machine systems to outpace the capabilities of foreign adversaries in years to come.[5] Indeed, official strategy documents and the formal statements of such leaders today hold as a given fact what military practitioners and scholars generally take years to realize – that a new technology is changing the character of human warfare itself.[6] The resultant expectation, at least according to some, is that underlying AI processes will lead to an inevitable transformation in the bases of national power and will alter the constitution of security relationships between states in both strategic and operational terms.

This article contributes to the nascent literature on AI and national security activities by outlining the ways in which AI is likely to alter the shape of, and strategic calculations bound up in, interstate cyber conflict.[7] While there is a small-but-growing body of work on the potential of AI for affecting military and national power writ large, surprisingly few reports exist that attempt to problematize AI in the context of state

5 The "Third Offset" is a strategy intended to be used by the Department of Defense in the United States to counter and overcome advances being made by key peer competitors, such as China and Russia, in areas of military modernization and technology development. The term "Third Offset" refers to previous efforts to overcome perceived positional, military or technological advantages held by the Soviet Union during the Cold War, the first of which originated with the famed Project Solarium convened by President Dwight Eisenhower in the 1950s. Robert Work, "Remarks by Deputy Secretary Work on Third Offset Strategy," Brussels, Belgium, April 28, 2016, accessed 1 February 2018, https://www.defense.gov/News/Speeches/Speech-View/Article/753482/remarks-by-d%20eputy-secretary-work-on-third-offset-strategy/; Cheryl Pellerin, "Deputy Secretary: Third Offset Strategy Bolsters America's Military Deterrence," *DOD News* October 31, 2018, accessed 1 February 2018: https://www.defense.gov/News/Article/Article/991434/deputy-secretary-third-offset-strategy-bolsters-americas-military-deterrence/; and Katie Lange, "3rd Offset Strategy 101: What It Is, What the Tech Focuses Are," *DODLive* March 30, 2016, accessed 1 February 2018, http://www.dodlive.mil/2016/03/30/3rd-offset-strategy-101-what-it-is-what-the-tech-focuses-are/.

6 This point refers to the oft-cited manifestation of revolutions in military affairs (RMA) that dot human history. On the historical emergence of the RMA, see Dima Adamsky, *The Culture of Military Innovation: The Impact of Cultural Factors on the Revolution in Military Affairs in Russia, the US, and Israel* (Redwood City, CA: Stanford University Press, 2010) and Benjamin Jensen, "The Role of Ideas in Defense Planning: Revisiting the Revolution in Military Affairs," *Defence Studies*, forthcoming. On the distinction between a revolution in military affairs and military revolutions more broadly, see MacGregor Knox and Williamson Murray, eds., *The Dynamics of Military Revolution 1300-2050* (Cambridge: Cambridge University Press, 2001).

7 For a broad overview of the scope and dynamics of cyber conflict, see inter alia Brandon Valeriano and Ryan C. Maness, *Cyber War Versus Cyber Realities: Cyber Conflict in the International System* (Oxford University Press, USA, 2015); and Christopher Whyte and Brian Mazanec, *Understanding Cyber Warfare: Politics, Policy and Strategy* (Oxon and New York: Routledge, 2018).

competition online.[8] Moreover, what work does exist tends to involve only descriptive analysis of threat scenarios, pulling up short of considering how AI's augmentation of cyber capabilities – specifically the application of machine learning techniques to attack and defense – alters the dynamics of strategic engagement in the digital domain.[9] This article aims to act as a resource for those interested in thinking more clearly about how AI stands to alter the dynamics of both interstate conflict processes and cyber conflict processes more specifically.

In the sections below, I illustrate how AI-driven cyber attacks differ dramatically in their form from conventional digital threats. I then argue that, although such forms of attack are possible and likely beyond the digital domain, the centrality of cyberspace to the deployment and operation of soon-to-be-ubiquitous AI systems implies new motivations for operation within the domain. This dynamic, alongside the prospect of cyber offense and defense upgraded by AI, challenges several assumptions held by current strategies for cyber conflict prevention and should be a cause of significant concern for policymakers.

I proceed in three sections. First, I address the task of defining artificial intelligence as it is relevant to cyber operations. Here, I highlight the manner in which machine learning – technically a subfield of AI research that, according to many, now virtually demands consideration as its own technology – promises to affect many of the assumptions about operations in cyberspace that have been considered as standard among security practitioners and researchers for many years. I then describe the practical advancements to be expected with AI-driven cyber operations, as distinct from those that more substantially depend on the hacker in the loop, and categorize two particular forms of AI cyber attack. I then engage the topic of recent cyber conflict strategy and discuss AI developments in context, before concluding.

8 For the limited work to date on AI and strategic studies, see inter alia Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence," *International Studies Review* (2019); Joe Burton and Simona R. Soare, "Understanding the Strategic Implications of the Weaponization of Artificial Intelligence," in *2019 11th International Conference on Cyber Conflict (CyCon)*, vol. 900, 1-17 (IEEE, 2019); Kareem Ayoub and Kenneth Payne, "Strategy in the Age of Artificial Intelligence," *Journal of Strategic Studies* 39, no. 5-6 (2016): 793-819; Heather Roff, *Advancing Human Security Through Artificial Intelligence*, (Chatham House, May 2017, https://www.chathamhouse.org/publication/advancing-human-security-through-artificial-intelligence; Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review* (2018); Kenneth Payne, *Strategy, Evolution, and War: From Apes to Artificial Intelligence*, (Georgetown University Press, 2018); Heather Roff, "COMPASS: A New AI-Driven Situational Awareness Tool for the Pentagon?" *Bulletin of the Atomic Scientists*, May 10, 2018, https:// thebulletin.org/compass-new-ai-driven-situational-awareness-tool-pentagon11816; Kenneth Payne, "Artificial Intelligence: A Revolution in Strategic Affairs?" *Survival 60*, no. 5 (2018): 7-32; Michael C. Horowitz, Gregory C. Allen, Elsa B. Kania, and Paul Scharre, "Strategic Competition in an Era of Artificial Intelligence," *Center for a New American Security* (2018); Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. "The Malicious use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *arXiv preprint arXiv:1802.07228* (2018).

9 See, for instance, Enn Tyugu, "Artificial Intelligence in Cyber Defense," in *2011 3rd International Conference on Cyber Conflict*, 1-11 (IEEE, 2011); or Mariarosaria Taddeo and Luciano Floridi, "Regulate Artificial Intelligence to Avert Cyber Arms Race," *Nature* 556 (2018): 296-298.

2. ARTIFICIAL INTELLIGENCE IN THE AGE OF CYBER CONFLICT

The label 'artificial intelligence' denotes a basket of technologies whose common attribute is the capability (or a set of capabilities) to simulate human cognition, particularly the ability of the human brain to adaptively reason, learn and autonomously undertake appropriate actions in response to a given environment.[10] In an even broader sense than is the case with all things "cyber," AI encompasses an immensely diverse landscape of technologies and areas of scientific development, from computer science to mathematics and neuroscience. The utilization of "AI" as a descriptor by many studies to describe new capabilities invariably risks, at least on some level, misleading readers by implying that artificial intelligence is best thought of as a relatively monolithic underlying technology whose design features will define future conflict. In reality, the implications of AI are best thought of in terms of unique interactions that will inevitably occur as an incredible array of potential smart machine systems is plugged into extant societal processes. This section attempts to contextualize the diverse form of what many simply generically refer to as "AI" and considers the implications for new techniques for the conduct of cyber conflict.

A. Machines That Reason, Learn and Act Autonomously

Machine cognition, which today substantially enables the function of most industrial sectors in advanced economies, has been a topic of significant interest to scientists and philosophers for the better part of two centuries. From Charles Babbage and Ada Lovelace to Alan Turing,[11] many of the greatest minds of the post-Industrial Revolution era have made their names by advancing societal thinking on the possibility of machines that can mimic how humans behave, move and think. More recently, the modern field of artificial intelligence – a term that emerged only in the early latter half of the 20th century among cybernetics and computer engineering researchers[12] – has its roots as a discipline in the substantial post-war work of minds like Marvin Minsky, Norbert Wiener and John von Neumann, who asked if, given the context of recent advances in computing, a machine might be made that could realistically simulate the higher functions of the human mind.[13] For such researchers, the challenge of machine intelligence lay in moving beyond the mere programmability of emerging computer constructs to building complex thinking systems capable of concept formation,

10 Jensen et al.(n 8) 10.

11 For contemporary description of such efforts, see inter alia Alan Turing, "Computing Machinery and Intelligence," *Mind* 49 (1950): 433-60; John von Neumann, *The Computer and the Brain*, (New Haven: Yale University Press, 1958); Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements* (New York: Cambridge University Press, 2010); and Herbert Simon, "Artificial Intelligence: An Empirical Science," *Artificial Intelligence* 77, no. 2 (1995): 95-127.

12 Randolph Kline, "Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence," *IEEE Annals of the History of Computing* 33, no. 4 (October-December 2011): 5-16.

13 See Kline, ibid.; J. Moor, "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years," *AI Magazine* 27, no. 4 (2006): 87–91; and Bruce Buchanan, "A (Very) Brief History of AI," *AI Magazine* 26 (Winter 2005): 4.

environment recognition, abstract reasoning and self-improvement.[14] In the decades that have followed, of course, not only have such systems become commonplace in application to narrowly-defined societal functions, but competing schools of thought variously hold – for mathematical, neurological, evolutionary or computational reasons – that the future will see general learners whose ability to autonomously operate in the world matches and surpasses that of humans.

Today, AI, as applied broadly across areas of global society, is what researchers label "narrow" AI – not the "general" systems that are the focus of science fiction classics like *Terminator* or *I, Robot*, but limited applications of machine intelligence to discrete tasks.[15] Generally, though there is some crossover and some meaningful within-category differentiation, the technologies of AI might be thought of as existing across three main categories – (1) sensing and perception, (2) movement and (3) machine reasoning and learning.[16] Of these, the last is by far the one that is arguably most synonymous with AI as it is often portrayed in popular settings. In this category are a range of advances that encompass machines' abilities to interpret data, represent knowledge and understand information imbued with social meaning. By far the most significant area within this category is that of machine learning, the scientific study and development of approaches to pattern recognition and knowledge construction absent pre-programmed instructions on how to interpret data.[17] Machine learning is relatively simple to understand. Whereas conventional computing might involve the input of data to a non-learning algorithm in order to output some functional result, machine learning involves the input of both data *and* a desired result to an algorithm that infers, learns about a given issue represented in the data and then outputs another algorithm tailored to allow for intelligent engagement therewith.[18] In short, today's sophisticated AI techniques do not overwhelm computational challenges via the application of processing power so much as they more effectively study data to design a better process. In this way, AI promises to solve a traditional challenge in continuing to realize the promise of computers for human society – that the development of complex software to run on increasingly sophisticated systems means ever-growing demands on computer memory (both in storage and processing terms) and manifestations of human error in programming at scale. Machine learning compensates, not by building a better computer or catching those errors, but by allowing computers to sidestep such issues by programming and reprogramming themselves more efficiently.

14 McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. "A proposal for the Dartmouth Summer Research Project on Artificial Antelligence, August 31, 1955." *AI Magazine* 27, no. 4 (2006): 12-12; available at http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

15 Burton and Soare (n 8) 5.

16 Jensen et al.(n 8).

17 For an overview of machine learning, see Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning," *Nature* 521 (2015): 436-44. Also see V. Mnih et al., "Human-Level Control through Deep Reinforcement Learning," *Nature* 518 (2015): 529-33; and David Silver et al. "Mastering the Game of Go without Human Knowledge," *Nature* 550 (2017): 354-9.

18 For perhaps the most accessible description of machine learning at the point of operation, see Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, (Philadelphia, PA: Basic Books, 2015).

While machine learning involves these new processes and techniques for the direct mimicry of human cognition, the first two categories above include the technologies that are needed in order to allow machines to effectively move beyond internal processes to survey and operate within an environment. To some degree, of course, better sensing and perception are part and parcel of building better machine reasoning and learning algorithms. After all, effective mimicry of human cognition requires that such algorithms are able to interpret data and make inferences as a human might.[19] This involves an ability to consider language usage as a human might – i.e. more effective natural language processing (NLP)[20] – and a capability to construct and represent knowledge via ontological treatment. In this way, learner algorithms are able to move beyond simplistic statistical treatment of input data to identify concepts and connections that are sociological in nature.

Beyond the syntactic foundations of such advances in perception, however, much AI involves the development of new sensor systems that create data for algorithms to consume. Advances in camera systems and microwave sensors that allow for sophisticated text and imagery recognition via visual feeds, for instance, are critical to the function of new software that helps law enforcement more rapidly assess patterns in criminal behavior or traffic flow. At the same time, AI involves the construction of robotic systems that can more effectively gather data and can act as autonomous agents with the help of advanced learning software.[21] Though these areas of AI are less relevant for the discussion of cyber conflict in this paper, I address them further below.

B. Cyber Offense Enabled by AI

How might artificial intelligence augment or upgrade offensive cyber operations (OCO)? The conventional answer to such a question is simply that AI stands to make cyber attacks more powerful, to reduce the effectiveness of conventional defensive measures and to make powerful attacks more accessible for the median malicious online actor. More specifically, four prospective dynamics surrounding AI-enabled cyber offense seem worthy of note.

19 For a seminal description of perception as a component element of broader attempts to build deep learning and reasoning systems, see Nicola Jones, "The Learning Machines," *Nature* 505 (2014): 146-8.

20 For further information on NLP, see inter alia Stephen Deagelis, "The Growing Importance of Natural Language Processing," WIRED Magazine, February 2014, found at https://www.wired.com/insights/2014/02/growing-importance-natural-language-processing/; and Erik Cambria and Bebo White, "Jumping NLP Curves: A Review of Natural Language Processing Research," *IEEE Computational Intelligence Magazine* 9, no. 2 (May 2014): 48-57.

21 For further reading on intelligent machine vehicle systems, see *inter alia* Mario Gerla, Eun-Kyu Lee, Giovanni Pau, and Uichin Lee, "Internet of Vehicles: From Intelligent Grid to Autonomous Cars and Vehicular Clouds," *IEEE* (2014); and Alberto Broggi, Alex Zelinksy, Umit Ozguner, Christian Laugier, "Intelligent Vehicles," in *Springer Handbook of Robotics*, ed. B. Siciliano and O. Khatib, (Berlin, Heidelberg: Springer, 2016), 1627-56.

1) Attack Surface Analysis at Scale and Speed

First, AI programming portends a significantly increased threat to prospective cyber attack victims insofar as it enables analysis of the attack surface of targeted systems and victim entities at scale. This manifests at two levels. The first of these is the opportunity for malware to utilize incoming data obtained via infection of machines to probabilistically judge where and when further infection is likely to lead to some value return. An example of how such future AI-enabled malware might work can be found in the financial institution-targeting Trickbot malware encountered in just the past two years.[22] At the point of initial compromise, Trickbot functioned similarly to other worm-enabled malware seen since the mid-2010s. Once a foothold was established, however, multiple additional machines were compromised within minutes, without a clear pattern of target selection. Not only was the malware able to scale its attack at some speed; it also selected victims based on a "smart" analysis of prospective success in further infection. I place the word "smart" in quotation marks here because the malware was not truly utilizing the AI techniques baked into malware that many experts herald as coming in the near future, but rather was manually programmed to take more careful action. Nevertheless, the example stands as a case wherein rapid understanding of the attack surface of a target network led to an unusual strategy of infection – not every potential target was hit, only those with clear vulnerabilities in the form of outdated Server Message Block (SMB) services – that proved difficult and costly for defenders set up to handle less persistent threats.

The second manifestation of greater analysis of attack surfaces leading to increased digital insecurity lies in the wealth of data and metadata that either might be obtained via traditional intelligence methods or are already available from criminal sources. The more data available to malicious actors interested in leveraging the advantages of AI for cyber aggression, the more capable the techniques employed might be. The future may very well hold cyber campaigns of either a criminal or a political nature which would be substantially informed by the wealth of data that might be made available to attackers for analysis. The gold standard of AI-enabled OCO, particularly those that target broad populations or large institutions, is one substantially designed by learning systems that infer lateral approaches to targets – and, in some cases, rapidly and autonomously undertake malicious action informed by such inference – with relatively low risk of detection or mitigation. Indeed, this threat of attack surfaces under sophisticated machine intelligence scrutiny is one of the core challenges that promises to impact current thinking on cyber conflict strategy and signaling. I return to this point in detail below.

22 For a description of the episode in context, see Cyber-Attacks, AI-Driven. "The Next Paradigm Shift." Also see Lior Keshet, "An Aggressive Launch: TrickBot Trojan Rises with Redirection Attacks in the UK," *Security Intelligence* (2016); and Darrel Rendell, "Understanding the Evolution of Malware," *Computer Fraud & Security* 2019, no. 1 (2019): 17-19.

2) Technique Adaptation

A second dynamic surrounding AI-enabled cyber offense is the ability of malware to autonomously select from a toolkit of options for further spread. Malware that is inserted into a machine might undertake environmental analyses and determine that another technique is more suited to attaching new victims than was the particular exploit involved in the initial compromise. Here, the shape of the AI-enabled cyber attack is not very different from the sophisticated software often employed by state security institutions or other advanced persistent threat actors (APTs). It is simply a more accessible, automatable method for empowering hackers of all stripes to utilize tools smart enough to fit variable elements of an attack toolkit into a diverse attack surface.

3) Adversarial Tactical Adaptation

Third, the threat of cyber offense upgraded by AI is also a type of malware that is able to adjust its own strategy of approach as operations are underway. Different from having a simple ability to assess potential targets and select appropriate methods of approach, AI programming will allow malware to alter its tactics in line with mission parameters as it learns more and more about the environment in which it is operating – and the defenders and users that populate that environment. Faced with diverse defense efforts across a diverse multi-network attack surface, a sophisticated AI-enabled attack on defense infrastructure could, for instance, determine that the rapid promulgation most advisable for one institution – say, a research laboratory – would be associated with greater risks of detection if executed against another target – say, a military base of operations. In such circumstances, the same piece of malware might be able to select an alternative approach, such as hiding or going "slow-and-low" in its effort to compromise machines and exfiltrate information. In this way, AI-enabled malware presents as an adversarial threat that functions even when – indeed, arguably especially when – robust defender efforts are apparent.

4) Multiple Mindsets

Finally, experts are concerned that AI-enabled malware will be able to analyze victim networks at scale and act autonomously to attack in ways that maximize opportunities for further compromise. A sub-element of the ability of AI-enabled malware to change tactical approach even beyond the point of victim identification and promulgation is the opportunity for multi-purpose malware that might change its own task or learn new tasks within the context of an existing operation. AI programming will allow sophisticated malware to learn about the defensive environment and compartmentalize lessons learned, such that alternative "mindsets" can drive activity where mission parameters are deemed to have changed (for example, upon discovery of a supervisory control system or where information has been retrieved and the task becomes one of exfiltration).

C. Cyber Artificial Intelligence Attacks: Threat Types

Naturally, if the potential underlying artificial intelligence for cyber offense can be summed up as greater adaptability, rapidity and opportunity for unexpected malicious behavior, then something similar can be said of the potential of AI-enabled cyber defenses. And indeed, it would be unfair to broach any discussion of the prospective impact of AI on cyber conflict without considering that the new learning, reasoning and sensing techniques will also come to – and already have begun to – undergird the efforts of defenders. Just as AI stands to augment and enhance the offense, so too will it become a necessity for those humans in the loop whose conventional perimeter, simulative and dissimulative defenses become the fodder from which adversarial attack AI builds better offensive routines.[23] Even here, however, it would be disingenuous to suggest that the AI arms race in cyber capabilities can be boiled down to tit-for-tat improvements in the relative capacities of those on the offense or defense. There are complex challenges facing those on the defense in the form of cyber artificial intelligence attacks (CAIA), which are attacks that seek to take advantage of approaches to system operations and defender routines in practice in order to subvert the legitimate functionality thereof.[24] In other words, CAIA essentially constitutes attacks against the AI itself that will increasingly come to underwrite cyber conflict processes. Such attacks might fall into two categories.

1) Input Attacks

Input attacks are those forms of contestation that seek to fundamentally mislead an AI system and skew the efforts of that system to classify patterns of activity.[25] If the expectations of a model designed by a learning AI program can be subverted, new space opens for unique, hard-to-predict exploits. Notably, input attacks do not involve attacking the code of AI systems or plugins itself; rather, the point of input attacks is deception that aims to control – or, at least, partially shape – how an AI system is "thinking" about a given issue or functional challenge. In this way, input attacks are best thought of as counter-command and control (counter-C2) warfare.[26]

Input attacks are highly varied in their form and can functionally be a great many things. This is because input attacks are defined by the function and deployment of the models they target. They might even involve physical activities in aid of cyber outcomes. For instance, a hypothetical re-running of the Stuxnet attack on Iran's

23 For discussion of simulation as an element of strategic interactions in cyberspace, see Erik Gartzke, and Jon R. Lindsay, "Weaving Tangled Webs: Offense, Defense, and Deception in Cyberspace," *Security Studies* 24, no. 2 (2015): 316-48.

24 The term "cyber artificial intelligence attacks" is inspired by its recent usage in Marcus Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It*, Belfer Center for Science and International Affairs, Harvard Kennedy School, August 2019.

25 Ibid. 19.

26 See Norman B. Hutcherson, *Command and Control Warfare. Putting Another Tool in the War-Fighter's Data Base*. No. AU-ARI-94-1, (Air Univ Maxwell AFB AL Airpower Research Inst, 1994); and Jeffrey A. Harley, *The Role of Information Warfare: Truth and Myths*, (Naval War Coll Newport RI Joint Military Operations Dept, 1996).

uranium enrichment facility at Natanz in which the defenders employed AI in the defense of internal networks might necessitate a nascent phase wherein the malware lay dormant vis-à-vis its core purpose and undertook secondary actions to install internal methods of subverting key defender system functions. At the same time, however, the malware might also benefit from input attacks undertaken by human intelligence assets. For instance, a piece of tape placed on one or more computer monitors on-site could conceivably trick security cameras into believing that those monitors were always on. Those cameras would not then flag an anomaly when malware turned a machine on during a period of inactivity.

2) Poisoning Attacks

In contrast with input attacks, poisoning attacks are activities that fundamentally seek to compromise the AI programming employed in enemy systems.[27] In the Stuxnet *redux* example above, such an attack on the part of the malware involved might, among other things, involve gradually increasing traffic volume to certain machines during non-peak hours. Therein lies the primary way in which AI systems are "poisoned" – the manipulation of data that such systems are trained upon so that the model learned by the target system does not accurately reflect reality. In poisoning an AI system, attackers in essence create backdoors via which further offensive action might be taken. This can, naturally, take a number of formats. An attacker might "train" a defending model to be oblivious to specific forms of anomalous behavior. Likewise, a system might be persuaded to fail or trigger some otherwise unrelated – but useful – process at a particular time when a certain action, such as a diagnostic scan, is taken.

3) Thinking About CAIA at Scale

It is tempting to primarily think of AI-enabled attacks as targeting the functionality of AI systems which defenders increasingly rely on to undertake security actions. However, the implications of CAIA for national security apparatuses go beyond such considerations. Specifically, the problem of poison for modern security institutions exists in such a way that the cyber-specific context implied in the threat type descriptions above constitutes only one element of the challenge. Given the coming proliferation of AI across military functions, security planners face the threat of skewing from nigh-uncountable sources. If adversary militaries wish to skew North Atlantic Treaty Organization (NATO) analytics, they might utilize conventional military deception methods – such as deploying decoy vehicles during military maneuvers to mislead NATO forces about the normal scale and dispersion of adversary forces – to do so as easily as they might tamper with training data via cyber means. Thus, it would be at least partially disingenuous to argue here that the augmentation of cyber conflict processes by AI constitutes a unique-to-the-domain coming transformation.

27 See Comiter (n 24) 28.

3. SHAPING BEHAVIOR IN AN AGE OF ADVERSARIAL LEARNING[28]

What *is* particularly unique about the intersection of artificial intelligence and cyber conflict processes, however, is that the centrality of cyberspace to the deployment and operation of soon-to-be-ubiquitous AI systems implies new motivations for operation within the domain. The prospect of subverting AI-driven security functions – in particular, the prospect of fundamentally poisoning the deliberative and operational bases of important national security establishment functions – provides incentive for operation in cyberspace beyond in-domain effects and outcomes. On the one hand, cybersecurity experts might expect an intensification of cyber conflict and criminal activities around the world based on near-term adoption of advancing AI programming that promises rapid adaptability and sophistication without either major investment or the need for major human presence in the loop. On the other hand, the same experts might expect an intensification of such activities because CAIA will so clearly often involve effects beyond the domain (e.g. cyber operations that are not operationally focused on some digital compromise so much as they are intended to affect real-world approaches to risk management, strategic assessment and resultant military deployments, financial outlays, etc.).

In the remaining section of this paper, I consider the implications of AI-augmented cyber attacks and CAIA for current strategic approaches to the mitigation of cyber conflict. Specifically, I describe the strategy of forward defense based around the dynamics of persistent engagement between adversaries in the domain that now constitutes American Title 10 approaches to operation online and suggest several core problems that either intensify or newly manifest in an era of large-scale proliferation of AI in cyber. The focus on U.S. strategy is intentional, as changes to America's force posture in the fifth domain represent the concrete edge of efforts to adapt prevailing approaches to cyber conflict in the context of both intensifying digital interference since 2010 and the failing applicability of legacy security concepts to the challenge. The dynamics of AI-augmented cyber conflict and the related questions that must be addressed vary beyond the scope of such singular focus, of course. But national contextualization allows for more in-depth exploration and produces analytic outcomes generalizable beyond the specific case.

A. Persistent Engagement and Defending Forward

In 2018, as it was elevated to the status of unified combatant command within the U.S. military, Cyber Command promulgated a new strategic vision centered around the

28 The phrase "adversarial learning" is a common one utilized by computer scientists to describe how machine learning algorithms are capable of adapting to hostile operational environments by crystalizing alternative – rather than combative – approaches to operation. See *inter alia* Daniel Lowd and Christopher Meek, "Adversarial Learning," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ACM (2005): 641-47; and Pavel Laskov and Richard Lippmann, "Machine Learning in Adversarial Environments," *Machine Learning* 81 (2010): 115-19.

concept of persistent engagement.[29] To put the concept and the strategy that emerges therefrom bluntly, persistent engagement means that Cyber Command intends to be everywhere, constantly maintaining presence and employing necessary tools against America's adversaries in networks wherever they might be found. The strategy pushes back against strategy as practiced in the past by both American administrations and allies, wherein operations were based on the political desire to mitigate cyber risk principally via norm development and through deterrent efforts that stemmed substantially from Cold War postures.[30]

In terms of the strategic logic of engagement in the domain, the persistent engagement strategy largely emerges from the work of Harknett and Fischerkeller in their time as scholars attached to Cyber Command. The authors argued that the unique character of cyberspace means that traditional deterrent approaches are doomed to failure.[31] Given that deterrence involves strong demonstrations of defense or meaningful statements of punishment following attacks, the prospects for developing a sustainable deterrent posture online are limited.[32] It is extremely difficult to demonstrate defensive capabilities at the scale demanded by a national cyber deterrent strategy, and punishment rarely works in the way it is intended. Communicating specific meaning in retaliation is difficult, particularly where the diversity of activities that constitute cyber conflict is immensely high. Moreover, response options are often not ready to go in the timeframe required by policymakers who seek to deter. Further, conceptual agreement on the significance or role of certain elements of the domain is not easy to come by, with poor understanding of what might be meant – if anything – by sovereignty online being a hallmark of the digital world.

The result is an alternative strategy – persistent engagement – that emphasizes "defending forward." This posture involves cyber forces operating beyond government and domestic networks to actively contest enemy activities aimed at harming national security or other national interests. Such operations, it is argued, can avoid escalation by embracing the doctrine of selective engagement and can be designed specifically to scale tactical efforts into strategic gains. In doing so, the idea is that the behavior of adversaries can be shaped and the scope of what is deemed to be appropriate

29 Department of Defense, *National Cyber Strategy of the United States of America*, 2018.

30 Nakasone, Paul M. "An Interview with Paul M. Nakasone," *Joint Forces Quarterly* (2019). https://ndupress.ndu.edu/Portals/68/Documents/jfq/jfq-92/jfq-92_4-9_Nakasone-Interview.pdf.

31 Fischerkeller, Michael P., and Richard J. Harknett. "Deterrence is not a credible strategy for cyberspace." *Orbis* 61, no. 3 (2017): 381-393.

32 For the broad literature on deterrence in cyberspace, see *inter alia* Libicki, Martin C. *Cyberdeterrence and cyberwar*. (Rand Corporation, 2009); Lupovici, Amir. "Cyber warfare and deterrence: trends and challenges in research." *Military and Strategic Affairs* 3, no. 3 (2011): 49-62; Crosston, Matthew D. "World gone cyber MAD: How 'mutually assured debilitation' is the best hope for cyber deterrence," *Strategic Studies Quarterly* 5, no. 1 (2011): 100-116; Jensen, Eric Talbot. "Cyber deterrence." *Emory Int'l L. Rev*. 26 (2012): 773; Denning, Dorothy E. "Rethinking the cyber domain and deterrence" (2015); Iasiello, Emilio. "Is cyber deterrence an illusory course of action?" *Journal of Strategic Security* 7, no. 1 (2014): 54-67; and Tor, Uri. "'Cumulative Deterrence' as a New Paradigm for Cyber Deterrence," *Journal of Strategic Studies* 40, no. 1-2 (2017): 92-117.

competition can be made known.[33] The resultant condition should, it is hoped, be one of "agreed competition" wherein the bounds of cyber conflict that are deemed to be acceptable can be consistently made known and where the worst excesses of digital insecurity for states might be avoided by the institution of precise conditions of case-by-case deterrence.[34]

B. Basic Challenges of AI for Persistent Engagement

Thinking effectively about the problem of poison for cyber conflict processes – particularly as a subset of all national security processes – is tricky, in that we have to fundamentally think about learning as it manifests in two different settings: in the organizational setting and in the construction of AI systems. It is not simply enough to consider the impact of rapid learning techniques for cyber conflict as we understand it today, though that approach to thinking about the problem of AI in this area *does* suggest some obvious challenges to be faced by prevailing strategy.

Above almost all other implications, broad-scoped upgrading of "conventional" cyber techniques portends a narrowing of the space within which adversaries might undertake cost-benefit calculations and come to believe that the benefits of further action are outweighed by the costs that might be imposed in the domain by forward defenders. Simply put, if smart tools exist that can more reliably avoid detection, take lateral routes to targets, or scale effects much more quickly than is the norm today, then adversaries are likely to exhibit increased willingness to continue operations under circumstances where they would not previously have done so. Especially given that the stakes of defection from agreed conditions of competition are not typically very high in political terms, this contraction of that space, wherein persuasion is argued to be possible under a doctrine of persistent engagement, ostensibly makes meaningful signaling even trickier from situation-to-situation. Likewise, at the most basic level, the proliferation of relatively robust abilities to achieve effects in the digital domain via lateral action – i.e. action that takes indirect, harder-to-predict pathways toward targets and outcomes – suggests that we might see recurrent incidents in areas where a threat had previously been thought to have been realized and countered in some form.[35]

33 Fischerkeller, Michael P., and Richard J. Harknett. "Persistent Engagement, Agreed Competition, Cyberspace Interaction Dynamics and Escalation." Orbis (Summer 2017) 61, no. 3 (2018): 381-393.

34 See *inter alia* Defense Science Board, Department of Defense. 2017. "Task Force on Cyber Deterrence." Defense Science Board, 3, 4. https://apps.dtic.mil/docs/citations/AD1028516; Bolton, John. 2018. "Transcript: White House Press Briefing on National Cyber Strategy - Sept. 20, 2018." Washington DC (September 8). Available at https://news.grabien.com/making-transcript-white-house-press-briefing-national-cyber-strategy.

35 This point references the oft-cited framing of cyber conflict history in the West as emerging via a series of realization episodes that have prompted a series of institutional and doctrinal adaptations over the past three decades. See Jason Healey (ed.), *A Fierce Domain: Conflict in Cyberspace, 1986 to 2012*, Cyber Conflict Studies Association, 2013.

It is perhaps most particularly worth noting that AI-enabled cyber conflict adds a new dimension to the traditional perception problem experienced in cyberspace, where attribution of intent or agency is particularly difficult at the point of threat detection and analysis.[36] Where a probing attack or some other action is detected, it is rare that the investigator is able to discern between run-of-the-mill adversary efforts to conduct espionage or some attacking action. In the near term, another possibility is that cyber actions may be not linked with either espionage or direct attack, but rather with attempts to interfere with the function of AI programming.[37] The particular danger here is that such attempts may involve activities that are even less clearly discernible as aggressive or not than is the case with espionage activities.

C. The Learning Problem

Beyond the basic challenges to the strategy of persistent engagement posed by the intensification of cyber conflict driven by the adaptability and rapidity brought by AI, of course, policymakers and practitioners must inevitably grapple with increasing uncertainty around the state of common knowledge between actors in the domain. The perception dynamic described above, for instance, is uniquely concerning for current strategic thinking on cyber conflict management, insofar as cyberspace is likely to be the domain of political activity most central to efforts to poison or otherwise interfere with AI systems. In a future where conflict involves broad-scoped efforts to manipulate the construction and operation of AI systems attached to myriad societal functions, cyberspace constitutes the primary highway via which such shaping efforts will likely flow. Moreover, state interest in operations of a poisoning nature via cyberspace is likely to grow over time as opportunities proliferate for the manipulation of processes that underlie strategy development, force posture determination and more.[38] Both of these points mean that strategic efforts to constrain adversaries' cyber actions relative to in-domain considerations may fail simply because they are not effectively armed with appropriate assumptions about the motivations of actors to operate online.

More broadly, the advent of narrow AI baked into most functional elements of a state's national security apparatus implies an enduring tension in the conduct of persistent operations intended to shape adversary behavior. All else being equal, the existence of robust AI systems on the part of foreign adversaries implies a learning problem –

36 See *inter alia* Nicholas Tsagourias, "Cyber Attacks, Self-Defence and the Problem of Attribution," *Journal of Conflict and Security Law* 17, no. 2 (2012): 229-44; Jon R. Lindsay, "Tipping the Scales: The Attribution Problem and the Feasibility of Deterrence Against Cyberattack," *Journal of Cybersecurity* 1, no. 1 (2015): 53-67; and Thomas Rid and Ben Buchanan, "Attributing Cyber Attacks," *Journal of Strategic Studies* 38, no. 1-2 (2015): 4-37.

37 This issue lies at the heart of what Buchanan labels the "cybersecurity dilemma." See Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust, and Fear Between Nations* (Oxford: Oxford University Press, 2016).

38 This assertion is quite arguably backed by work that demonstrates in both quantitative and qualitative terms an increasing turn towards political warfare as an adjunct of cyber conflict, in line with the proliferation of digital services and social platforms that undergird major societal functions. See, for instance, Brandon Valeriano, Benjamin M. Jensen, and Ryan C. Maness, *Cyber Strategy: The Evolving Character of Power and Coercion* (Oxford: Oxford University Press, 2018).

the more security institutions operate to shape behavior, the more those adversaries *should* be empowered to understand and overcome such strategies. After all, much as in the case of Generative Adversarial Networks (GANs) that study the actions of AI models in order to continually improve offensive capabilities,[39] AI-enabled cyber forces presented with unique patterns of behavior-shaping attack from abroad will naturally undergo a process of adversarial learning where foreign action does not bound the shape of acceptable behavior so much as define the criteria under which future aggression is probabilistically less likely to induce some cost. Particularly given the incentive described above towards the use of AI-enabled software agents that have dramatically higher track records of success – given their adaptability – than non-AI-enabled versions, the commonplace existence of such systems seems likely to work against the development of static norms of behavior.

Finally, the result of an emergent era in which AI-driven adversarial learning is the key feature of interstate interactions online is a perpetual challenge of validation. In recent scholarship, there have already been some discussions about the challenges involved in applying relevant metrics to the strategy of persistent engagement such that defense practitioners might determine its effectiveness.[40] Such challenges multiply, given the AI-ification of cyber conflict processes and the problem of poison as a regular feature of operation in the domain. Whereas analysis of broad patterns of activity might otherwise offer some indication as to the effectiveness of forward defensive efforts aimed at dissuading particular adversary behaviors, such metrics may not apply in a significant fashion in an era where counter-action from foreign peers is not expected to be tit-for-tat, but rather entirely alternative in approach. In other words, where the paradigm of operation shifts from in-kind engagement – even if that engagement emerges from an admittedly diverse toolkit – to an imperative of lateral approach and misdirection, attempts to validate current strategic processes seem likely to be ineffective beyond simplistic analysis of major event incidence.

4. IMPLICATIONS FOR STRATEGIC THINKING

The purpose of this article is to contribute to the nascent literature on AI and national security activities by outlining the ways in which AI is likely to alter the shape and strategic calculations bound up in interstate cyber conflict. It is hoped that the sections above can act as a resource for those interested in thinking more clearly about how AI stands to alter the dynamics of both interstate conflict processes and cyber conflict processes. Naturally, a substantial part of the effort made herein has been definitional. Indeed, it is from the categorization of different threat forms linked to the

39 Vincent, James. "Deepfake Detection Algorithms Will Never Be Enough." *The Verge* (2019).

40 See, for instance, Jason Healey and Neil Jenkins, "Rough-and-Ready: A Policy Framework to Determine if Cyber Deterrence is Working or Failing," in *2019 11th International Conference on Cyber Conflict (CyCon)*, vol. 900 (IEEE, 2019): 1-20.

augmentation of cyber conflict processes by AI models and systems that the primary argument of this paper emerges – that the centrality of cyberspace to the deployment and operation of soon-to-be-widespread AI systems implies new motivations for operation within the domain. The implications thereof for current cyber conflict strategies – particularly those being worked on by Western defense establishments – are numerous and remain to be assessed in full as literature on the subject is developed in the future. Nevertheless, some immediate takeaways are apparent.

First, strategic planners and policymakers must recognize from the start that there are two levels of challenge when it comes to AI augmentation of cyber conflict processes. At the first level, AI promises to reduce the window in which it may be possible to shape competition in cyberspace in favorable terms. At the second, AI intensifies and adds a new dimension to the challenges of validity and attribution already present in cyber operations. Simply put, given the opportunities for poisoning by soon-to-be-ubiquitous AI models at work in security apparatuses, how can defenders really know what they think it is they know about the integrity of their systems? At the strategic level, given that broad-scoped attempts to shape competition between AI-enabled adversaries are likely to empower opponents via a process of adversarial learning, how can policymakers and military practitioners really know what they think it is they know about strategic conditions?

Second, because of the various challenges bound up in effectively deploying AI for national security purposes, the effectiveness thereof is likely to be bound up in the approach organizations take to trusting their AI systems and to managing the interaction of human and machine operators.[41] Much of what has been discussed in the sections above involves – to at least some degree – the problem of ghosts in the machine, where it is human assumptions present in the code of machine intelligence systems that form the true problem for effective deployment for national security purposes. While such problems are arguably unavoidable as we move toward more common employment of AI than is the case today, it seems likely that protocols for keeping humans in the loop at critical junctures are part of the solution to problems of (either malicious or self-inflicted) poison.

Finally – and perhaps most significantly – it seems clear that, in the forthcoming era of AI-enabled contestation in world affairs, strategy development, assessment and validation must emerge significantly from cross-domain understanding of the strategic motivations of adversaries. If cyberspace is not only a domain wherein unique forms of contestation and signaling can take place, but is also the most significant terrain over which actions can be taken to affect processes that underlie all areas

41 This is not a thus-far uncommon argument made by scholars of cyber conflict. See, for instance, Rebecca Slayton. "What is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment," *International Security* 41, no. 3 (2017): 72-109.

of modern society, then strategic planners would do well to build from assumptions that move beyond simple logic-of-the-domain characterizations of digital affairs. As has previously been argued in both implicit and explicit terms,[42] cyber conflict so often manifests in aid of non-digital contestation that we would do well to couch our analyses in terms of the logic of conflict processes *other* than cyber. This stands to be especially the case with artificial intelligence, not least given the fact that the targeting of AI for security purposes is so likely to be significantly tied to use of the computer and Internet systems upon which such programming must inevitably run.

42 See, for instance, Christopher Whyte, "Dissecting the Digital World: A Review of the Construction and Constitution of Cyber Conflict Research," *International Studies Review* 20, no. 3 (2018): 520-32; and Jon R. Lindsay, "Stuxnet and the Limits of Cyber Warfare," *Security Studies* 22, no. 3 (2013): 365-404.