Poison, Persistence, and Cascade Effects: AI and Cyber Conflict

Christopher Whyte

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 nefarious AI techniques is particularly worrisome, not least because prototype applications already exist. Cyber attacks augmented by AI portend tailoring and manipulating the human side of important societal systems as well as introducing the risk that comes from moving technical skill from the hacker to an algorithm. The rise of AI-augmented cyber defenses incorporated into national defense postures will likely be vulnerable to "poisoning" attacks that predict, manipulate, and subvert the functionality of defensive algorithms. These AI-enabled cyber campaigns contain great potential for operational obfuscation and strategic misdirection. At the operational level, piggybacking onto routine activities to evade security protocols adds uncertainty, complicating cyber defense particularly where adversarial learning tools are employed offensively. Strategically, AI-enabled cyber operations may be able to pursue conflict outcomes beyond those expected of adversaries. Perhaps more worrisome is that the centrality of the Internet to new AI systems incorporated across all areas of national security—not just to cyber conflict processes—indicates that sophisticated adversaries may be motivated to launch offensive online actions to achieve effects in other domains with some increasing regularity.

*****

In recent decades, few technological developments have captured the attention and sparked the concern of national publics so much as those linked to artificial intelligence (AI). This might seem a remarkable and outlandish statement given that, if prompted, the average 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, been at the heart of prominent conversations about the future of human innovation and the changing shape of societal security.[1] Tech luminaries con-

tinue to expound the revolutionary potential of new machine learning and reasoning techniques that now easily solve endemic issues of overcomplexity 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 in an attempt to convince audiences of the stakes involved in getting AI "right."[3]

Around the world, few entities are as focused on the impact AI systems portend for security as are 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.[4] Indeed, official strategy documents and formal statements maintain something military practitioners and scholars generally take years to realize—that a new technology is changing the character of warfare itself.[5] The resultant expectation, according to some, is that underlying AI processes will lead to an inevitable transformation in the bases of national power and alter security relationships between states in both strategic and operational terms. While there is a small but growing body of work on the potential of AI to affect military and national power writ large, surprisingly few reports attempt to discuss AI developments in the context of state competition online.[6] Moreover, what work does exist tends to involve only descriptive analyses of threat scenarios, without considering how AI's augmentation of cyber capabilities—specifically the application of machine learning techniques to offense and defense—alters the dynamics of strategic engagement in the digital domain.[7]

AI-driven cyber attacks differ dramatically from the more conventional digital threats that have occupied practitioners and researchers for the past three decades. Their effects are also possible—even likely—to be felt outside of cyberspace. However, the centrality of cyberspace to the deployment and operation of soon-to-be ubiquitous AI systems implies new motivations for operations within the cyber domain. The prospect of offensive and defensive cyber operations upgraded by AI challenges several assumptions held by current strategies for cyber conflict prevention and should be a cause of significant concern for policy makers. AI is likely to alter the shape and strategic calculations bound up in interstate cyber conflict and alter the dynamics of interstate cyber conflict processes.[8] However, such transformation will not come simply from the sophistication of attack and defense by AI, but rather from the manner in which AI adds

new complexity and therein intensifies issues of strategic perception and misperception.

Ultimately, AI does not itself imply inevitable advantages for attackers over defenders (or vice versa). But adversarial learning techniques layer complexity on top of already complex operational conditions in cyberspace and may contribute to an uptick in offensive behavior. After all, nested logics of engagement across a heterogeneous global environment make for an even more convoluted battlespace than exists presently. Of greatest concern, however, is the centrality of the Internet to new AI systems that will be incorporated across all areas of national security, not just in cyber conflict processes. This inevitable application of new techniques and technologies across the national defense enterprise suggests that sophisticated adversaries may be motivated to launch offensive online actions to achieve effects in other domains with some increasing regularity. This introduces new challenges for defense at scale and amplifies some risks of AI-enabled engagements, such as the possibility of AI-driven "flash crashes."

This article takes steps to reconcile the task of defining artificial intelligence as it relates to cyber operations by highlighting how the major relevant area of AI development, machine learning, promises to affect many of the assumptions about operating in cyberspace that have been considered standard among security practitioners and researchers for some years.[9] Then, it categorizes the primary advances in AI technologies likely to augment offensive cyber operations, including the shape of cyber activities designed to target AI systems. Finally, the article frames the implications for deterrence in cyberspace by referring to the policy of persistent engagement (PE), agreed competition, and forward defense promulgated in 2018 by the United States.

Before moving forward, one clarification seems worthy of mention. This article is structured around a discussion of the utility of AI learning techniques for cyber offense. It does so as a basis for discussing the totality of strategic cyber considerations pertaining to AI. As implied above, however, it does not fundamentally argue that AI systematically favors the offense as some international relations scholars argue.[10] While new adversarial learning techniques do seem poised to enhance the attacker's toolkit over and above that of the defender, the logic of offense dominance with AI likely mirrors that of cyber operations: offense is dominant and tactical deterrence impossibly hard only where the value of target systems is high. Otherwise, AI stands to favor the defense as much as the offense, at least at the tactical level. This parity of effect may not bear out in the realm of

strategic interaction where new learning capabilities employed at scale in routine operations add complexity to the already murky perspective of operators who must consider interacting operational, institutional, and geopolitical contexts.

Artificial Intelligence and Assumptions on Cyber Operations

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.[11] 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. As such, using AI as a descriptor in many studies to describe new capabilities invariably risks, at least on some level, misleading readers by implying that AI is best thought of as a relatively monolithic underlying technology whose design features will define future conflict. 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 are plugged into extant societal processes. The challenge is to contextualize the diverse forms of what many generically refer to as AI and consider the implications of new techniques on the conduct of cyber conflict.

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, 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 mimic how humans behave, move, and think.[12] More recently, the modern field of *artificial intelligence*—a term that emerged only in the latter half of the twentieth century among cybernetics and computer engineering researchers—has its roots as a discipline in the substantial postwar work of AI pioneers like Marvin Minsky, Norbert Wiener, and John von Neumann.[13] They 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.[14] For such

researchers, the challenge of machine intelligence lay in moving beyond the mere programmability of emerging computer constructs to build complex thinking systems capable of concept formation, environment recognition, abstract reasoning, and self-improvement.[15] Such systems are now commonplace in application to narrowly defined societal functions. Moreover, 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 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 *The Terminator* or *I, Robot*, but limited applications of machine intelligence to discrete tasks.[16] Generally, though there is some crossover and 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.[17] Of these, by far the one most arguably synonymous with AI as it is often portrayed in popular settings is the last. In this category is a range of advances that encompass machine abilities to interpret data, represent knowledge, and understand information imbued with social meaning. By far the most significant area in this category is machine learning, the scientific study and development of approaches to pattern recognition and knowledge construction absent preprogrammed instructions on how to interpret data.[18] Machine learning is relatively simple to understand. We might think of conventional computing as involving the input of data to a (non-learning) algorithm that then outputs some functional result, such as a statistic or perhaps a graphical representation of the data. By contrast, machine learning involves the input of both data and a desired result to an algorithm (often called a "learner") that infers, learns about a given issue represented in the data, and then outputs another algorithm tailored to allow for intelligent engagement.[19] 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. Specifically, 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 manifestation of human error in programming at scale. Machine learning does not compensate by building a better computer or by just catching those errors more efficiently. Rather,

it does so by allowing computers to sidestep such issues entirely by programming and reprogramming themselves more efficiently.

While machine learning involves those new processes and techniques for the direct mimicry of human cognition, the first two categories above—sensing and perception and movement—include the technologies needed to allow machines to effectively move beyond internal process 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.[20] This involves an ability to consider language usage as a human might—that is, more effective natural language processing (NLP)—and a capability to construct and represent knowledge via ontological treatment.[21] Thus, learner algorithms can 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 act as autonomous agents with the help of advanced learning software.[22]

Expected Advances in AI-Enabled Cyber Offense

How might artificial intelligence augment or upgrade offensive cyber operations (OCO)? The conventional answer to such a question is simply that AI (specifically, machine learning) stands to (1) make cyber attacks more insidious, disruptive, and long-lasting; (2) reduce the effectiveness of conventional defensive measures; and (3) make powerful attacks more accessible for the median malicious online actor. Thus, AI portends unprecedented adaptability, rapidity, and opportunity for unexpected malicious behavior than has previously been the case. Four prospective dynamics surrounding AI-enabled cyber offense seem worthy of note.

Attack Surface Analysis at Scale and Speed

AI programming portends a heightened threat to prospective cyber-attack victims insofar as it enables analysis of the attack surface of targeted systems and victim entities at scale.[23] This manifests at two levels. The first is the opportunity for malware to use 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 comes from the financial sector–targeting Trickbot malware encountered in just the past two years.[24] At the point of initial compromise, Trickbot—the target of preemptive cyber operations conducted by Microsoft and US Cyber Command in October 2020 due to its prospective use in election interference activities—functions similarly to other worm-enabled malware seen since the mid-2010s. Once it establishes a foothold, however, within minutes the software targets and compromises additional machines that do not follow a clear pattern of target selection. Not only is the malware able to scale its attack at some speed, it also selects victims based on a "smart" analysis of prospective success in further infection. The word "smart" is placed in quotation marks here because the malware is not truly using the AI techniques that many experts herald as coming soon; rather, it is manually programmed to take more careful action. Nevertheless, the example stands as a case wherein a rapid understanding of the attack surface of a target network has led to an unusual strategy of infection. Not every potential target is hit but only, in the financial services case at least, targets with clear vulnerabilities in the form of outdated Server Message Block (SMB) services. The strategy there proved difficult and costly for defenders set up to handle less persistent threats.

Another 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 criminal or political natures that are 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 targeting 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 analysis is one of the core challenges that promises to impact current thinking on cyber conflict strategy and signaling.

Technique Adaptation

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

Adversarial Tactical Adaptation

The threat of cyber offense upgraded by AI is also one of malware able to adjust its own strategy of approach as operations are underway. Different from 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 operating environment and the defenders and users populating that environment. Faced with diverse defense efforts across a diverse multinetwork 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. Therefore, AI-enabled malware presents as an adversarial threat that functions even or especially when robust defender efforts are apparent.

Multiple Mindsets

Experts are concerned not only 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 multipurpose 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 (such as upon discovery of a supervisory control system or where information has been retrieved and the task becomes one of exfiltration).

Cyber Artificial Intelligence Attacks: Threat Types

Naturally, if the potential underlying AI for cyber offense can be summed up as greater adaptability, rapidity, and opportunity for unexpected malicious behavior, then something similar can be said for 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.[25] 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. Those on the defense face complex challenges in the form of cyber artificial intelligence attacks (CAIA), which seek to take advantage of approaches to system operations and defender routines in practice to subvert their legitimate functionality.[26] In other words, CAIAs essentially constitute attacks against the AI itself that will increasingly come to underwrite cyber conflict processes. Offense, then, becomes far more attractive to cyber-capable adversaries than it is currently because of the increased potential to achieve second-order effects (i.e., to affect more than just the targeted infrastructure with a single attack by manipulating underlying algorithmic behaviors). Such attacks might fall into two categories: input attacks and poisoning attacks.

Input Attacks

Input attacks are forms of contestation that seek to fundamentally mislead an AI system and skew its efforts to classify patterns of activity.[27] 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 plug-ins themselves. 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.[28]

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 those models they target. They might even involve physical activities in aid of cyber outcomes. For instance, a hypothetical rerunning of the Stuxnet attack on Iran's uranium enrichment facility at Natanz—wherein the defenders employed AI in the defense of internal networks—may have necessitated a nascent phase wherein the malware lay dormant vis-à-vis its core purpose. It would then undertake secondary actions to install internal methods of subverting key defender system functions. At the same time, the malware might also benefit from input attacks by human intelligence assets. For instance, a piece of tape placed on computer monitors on-site could conceivably trick security cameras into believing that those monitors are always on. Those cameras would not then flag an anomaly when malware turns a machine on during a period of inactivity.

Poisoning Attacks

In contrast with input attacks, poisoning attacks are activities that fundamentally seek to compromise the AI programming employed in enemy systems.[29] In the Stuxnet redux example above, such an attack on the part of the malware involved might, among other things, entail gradually increasing traffic volume to certain machines during nonpeak hours. Therein lies the primary way AI systems are "poisoned"—the manipulation of data that such systems are trained on so that the model learned by the target system does not accurately reflect reality. In poisoning an AI system, attackers create backdoors through which further offensive action might be taken. This can, naturally, take several 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.

Though the subject of poisoning attacks may be reasonably new in the literature on cyber conflict and national security, design of and defense

against such activities have long been a focus within the machine-learning literature in computer science. It would be disingenuous to suggest here that the threat is insurmountable. While much work has consistently demonstrated the limited access and resources required to engage in poisoning attacks on neural networks, a few strategies seem promising for defense on several fronts.[30] Use of blockchain or watermarking techniques to "sign" data as safe to use, for instance, might prevent compromise even when access by malicious attackers is possible.[31] Statistical optimization techniques using only subsets of data sets also decreases reliance on entire data repositories and allows for self-analysis of data provenance.[32] Others have suggested a strategy of introducing controlled perturbations into data to dramatically reduce the effectiveness of poisoning efforts.[33] Nevertheless, these defensive efforts are vulnerable to many of the conditional vulnerabilities that characterize the best network defense techniques. For instance, the need to apply such defenses at scale clashes with the inevitable complexity of the global information technology landscape and conflicts with commercial interests in product development that emphasize proprietary solutions at speed over best security practices. Thus, poisoning attacks promise to be an increasingly prominent threat to smart systems into the future, particularly as they benefit from the use of self-learning techniques to compensate for defender efforts.

Thinking About Cyber AI Attacks at Scale

While it is tempting to think of the threat of attacks that compromise the function of AI systems that defenders must increasingly come to rely on only at the level of cyber operations themselves, the implications of CAIAs for national security apparatuses go beyond such considerations. Specifically, the problem of poison for modern security institutions exists beyond the implications for cyber conflict; indeed, cyber operations are just one element of the challenge. Given the coming proliferation of AI across military functions, security planners face the threat of skewness from nigh uncountable sources. If adversary militaries wish to skew North Atlantic Treaty Organization (NATO) analytics, they might use 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—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.

Shaping Behavior in an Age of Adversarial Learning

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 expanded motivations—such as an increased interest in using cyberspace to affect extra-domain technological processes—for operations 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—incentivizes operations 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 cyber AI attacks will clearly so often involve effects beyond the domain (e.g., cyber operations 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.).

Implications for Deterrence in Cyberspace

What follows is a contextual analysis of the implications of AI-augmented cyber attack for current strategic approaches to mitigating cyber conflict. This includes the strategy of forward defense based around the dynamics of persistent engagement between adversaries in the cyber domain that now constitutes US Title 10 approaches to operations online. It suggests several core problems that either intensify or newly manifest in an era of large-scale proliferation of AI in cyber. The focus on US strategy is intentional; 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. Dynamics of AI-augmented cyber conflict and the ensuing 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 case.

Defending Forward and Persistent Engagement

In 2018, as it was elevated to the status of unified combatant command in the US military, Cyber Command promulgated a new strategic vision centered around the concept of persistent engagement.[34] To put the concept and strategy that emerge bluntly, PE means that Cyber Command intends be everywhere, constantly maintaining presence and employing necessary tools against US adversaries in networks wherever they might be found. The strategy pushes back against past practices by the US and its 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 the shape of Cold War postures.[35]

In terms of the strategic logic of engagement in the domain, the PE strategy largely emerges from the work of Richard Harknett and Michael Fischerkeller during their time as scholars attached to Cyber Command. The authors argue that the unique character of cyberspace means that traditional deterrent approaches are doomed to failure.[36] Given that deterrence involves strong demonstrations of defense or meaningful statements of punishment, they contend, prospects for developing a sustainable deterrent posture online are limited (or so the architects of the new approach hold).[37] 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 execute in the time frame required by policy makers that seek to deter. And 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 of Western nations 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 competition can be made known.[38] The resultant condition should, it is hoped, be one of "agreed competition" wherein the bounds of cyber conflict 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.[39]

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 difficult in that we fundamentally have to think about learning as it manifests in two different settings, 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 portend a simple functional challenge for cyber strategy. Specifically, it suggests 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 operating under circumstances they would not have previously. 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 yet more difficult 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—action that takes indirect, harder-to-predict pathways toward targets and outcomes—suggests that we might see recurrent incidents in areas where the threat had previously been thought to have been realized and countered in some form.[40]

It is worth noting on an operational level that AI-enabled cyber conflict adds a new dimension to the traditional perception problem experienced in cyberspace wherein attribution of intent or agency is particularly difficult at the point of threat detection and analysis.[41] 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 with attempts to interfere with the function of AI programming.[42] The particular danger here is that such attempts may involve activities even less clearly discernable as aggressive than is the case with espionage activities.

AI, Feedback Loops, and the Logic of Persistent Engagement

Beyond functional AI-induced issues of added sophistication and perception, the strategic logic of PE may be made more vulnerable when new learning tools employed at scale also impact second-order conditions relevant to the conduct of cyber conflict in broader international relations. Jason Healey, in his analysis of challenges awaiting the United States as it continues to commit to the strategy of persistent engagement, discusses such logic in the context of feedback loops.[43] Feedback loops describe any system where the outputs of a process either constitute or affect the inputs of that same process as it iterates over time. Positive and negative feedback mean, respectively, outcomes that either amplify the original process or dampen it. With PE, the idea is that forward operation allows the US to see attacks before they occur (informing domestic actors more effectively as a result) and produces "friction" that increases the costs of antagonism for adversaries.[44] Alongside more conventional deterrent operations, this activity should in theory create negative feedback—a dampening, constraining effect on aggressive behavior in cyberspace.[45]

In discussing PE in this fashion, Healey joins others concerned about the risks of such an assertive policy.[46] A main concern, what he refers to as "on-net" challenges, revolves around the issues of misperception and tacit intersubjectivity in direct cyber interactions discussed above.[47] Beyond simple functional difficulties, it is worthwhile reiterating in more detail that AI exacerbates a fundamental problem with PE as a strategy, namely that it includes no concrete method of communication other than conflict actions themselves. This particularly manifests on two fronts.

First, the assumptions of tacit bargaining as a critical pushback against the track record of deterrent efforts in cyberspace now functionally sit at the heart of American cyber conflict policy.[48] This is problematic because strategic assumptions must be based on a range of operational dynamics that are inevitably hard to fully observe from just one side of the screen. Friction designed to produce negative feedback is likely to fail if costs to adversaries are minimal.[49] Certainly, operators can design tactical actions to avoid such an outcome and maximize strategic gains.[50] But to some degree, the impact of forward defense efforts will always be a question of adversary infrastructure and resource commitment, about which the home team will

always have imperfect information. If reconstruction of infrastructure is inexpensive, friction will not work. Today, this is a concerning element of PE because the funding structures and priorities of authoritarian opponents can be relatively opaque. Likewise, a robust defense against PE aggression lies not only in in-domain actions but also in adversary efforts to build operational resilience. This may be the commitment of resources sufficient to regularly make American "friction" ineffectual at cost imposition. Or, somewhat more worrying, this might involve further decentralization of extensive cyber operations infrastructure on the part of adversaries, essentially adding distance and compartmentalization of assets with the use of internal, criminal, and non-state proxies to create redundancy and introduce obstacles to American efforts to map the battlespace.

Second, it is not fully clear what the "acceptable" behavior desired by the strategy of PE might look like.[51] As opposed to a strategy like that of the "fleet in being"—which some scholars have suggested as a more realistic strategic alternative to persistent engagement—that explicitly permits low-intensity antagonism, PE calls for setting norms of behavior to be defined by prevailing military and political stakeholders.[52] This means, as some have noted, that there may easily exist tactical or political reasons over time to attempt to interdict any aggressive behavior. And because the only communication intended under PE is in the method of engagement, mechanisms to quickly clarify expectations promise to be clunky at best.

Artificial intelligence adds to the challenges facing PE on both fronts. Currently, a major concern is that failed friction will lead to "aggression spirals" in which both sides escalate in search of costly digital territory. AI brings new dimensionality to this concern. Simplistically, AI is likely to lower costs of reconstruction of digital assets across the board, making this situation of failed friction more likely. After all, the game-changing fact of the revolution in machine learning amounts to an ability to overcome—via use of self-reprogrammable learner algorithms—the programming bloat that inevitably costs organizations resources as their infrastructure is called upon to provide more diverse specialty functions at scale. Additionally, in-domain escalations might be motivated beyond the link between offensive actions and imposed costs assumed by the strategy. Aggression spirals under controlled conditions—at least, as the adversary judges the risks and intentions involved—provide opportunities to train defensive platforms and to showcase strategies of aggression intended to mislead the peer competitor. Such activity is clearly attractive, as enough evidence of adversary behavioral preferences might create cognitive schema and

operational cultures that dampen tactical adaptability in the face of new patterns in the data.

The question of "acceptable" behavior also looms large given the question of AI in cyber. As laid out above, states are likely to be motivated to directly influence AI systems employed by adversaries, both those pertaining to cyber operations and those functionally at the heart of innumerable national security and societal processes. This dual focus on subversion of process and of process beyond domain-specific capacities makes answering the behavioral question even harder. If subversive attacks that have increasingly real meaning for strategic knowledge capabilities are imperative for competitors heavily invested in use of AI, then what conventional metric can possibly be used to gauge "aggression"? This is particularly salient given the way the PE strategy holds espionage apart as "acceptable" behavior. If low-intensity and lateral engagement begin to threaten core functional capabilities beyond what is currently the case, then strategists will be forced to either by demonstration or explicit declaration attempt to offer tighter definitions of what activity is "unacceptable" that parses apart espionage from poisoning operations. And such a development seems likely. After all, the logic of PE emerges in trusting that an invisible hand of "market" correction will work to produce behavioral equilibrium. The strategy would surely fail, at least in part, if trust in the integrity of that hand faltered. Actors must understand the limits of the game they are playing. The threat of a subverted rule set itself will likely motivate assertive action to stabilize the battlespace, adding yet another layer of complex calculation to daily action and reaction in the domain.

The issue of cyber conflict in an era where cyberspace is the primary highway for the operation of innumerable AI systems spread across important security and societal infrastructure bears additional mention in the context of PE. Forward defense is simply one layer of the US effort to limit aggression experienced via cyberspace.[53] Traditional deterrent operations and efforts to build norms using conventional diplomatic approaches remain as robust pillars of American cyber foreign policy. Persistent engagement is the lynchpin underlying these additional efforts (see fig. 1).

However, the success of PE seems likely only where there will be clear situational alignment with other efforts. In large part, this is because there is so much natural oscillation in the conditions of sophisticated cyber conflict actions and the reactions of complex state military and civilian government infrastructure. The context of much complexity in international affairs—including global and domestic politics, private versus public behavior in cyberspace, intelligence versus military use of the fifth domain, and

more—constructs nested spaces wherein contrasting perspectives about the logic of digital engagement make sense. Simply put, the "AI-ification" of advanced industrial states in the years to come is likely to cause the multiplication of such spaces as cyberspace becomes the central artery through which so much added manipulative traffic flows. This will make it harder for adversaries to be sensitive to each other's signals while at the same time motivating actions targeting non-domain effects as a strategy to degrade state confidence in the value of longitudinal data pertaining to cyber operations.

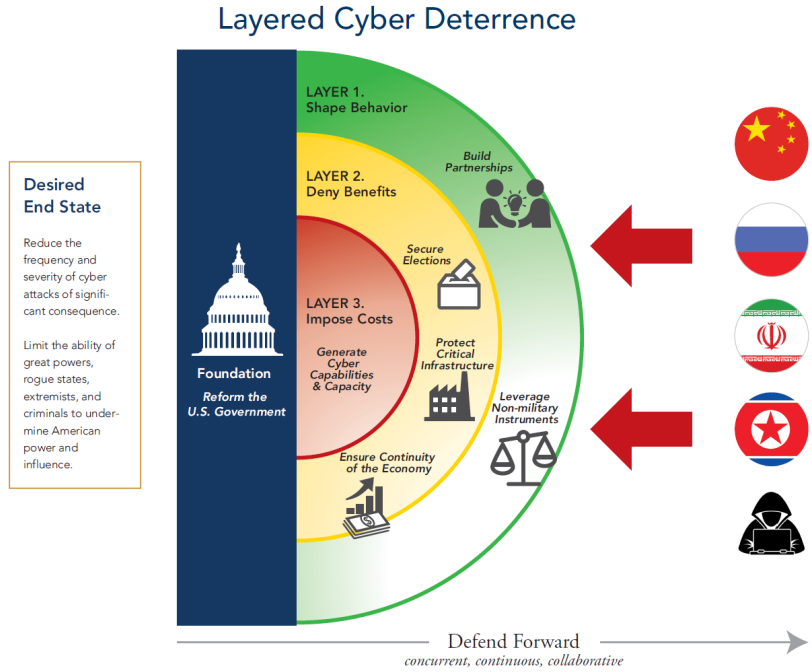

Figure 1. Layered Cyber Deterrence. (Reproduced from Angus King and Mike Gallagher, co-chairs, US Cyberspace Solarium Commission, *Cyberspace Solarium Commission Report* [Arlington, VA: US Cyberspace Solarium Commission, March 2020], 7, https://www.solarium.gov/.)

A final implication of AI for PE and current approaches to cyber conflict is with how efforts to secure cyberspace might degrade, as Healey notes, the reality of "an open, interoperable, reliable, and secure Internet that fosters efficiency, innovation, communication, and economic prosperity."[54] Forward defense naturally relies on a great deal of trust among allies, private sector partners in industry, and other elements of civil society. Yet the actions implied by the strategy are inevitably among the most invasive and assertive imaginable on the part of a national government like that of the United States. This is particularly the case given the way patterns of

engagement are unlikely to ever be the predictable intrusions of terrestrial conflict. This produces a trust challenge to the success of the strategy, with no easy solutions and many fault lines where irritation is not only possible but likely. Global outrage following leaks from Edward Snowden, the Shadow Brokers, and more was not limited to foreign states and persons but was also common in the American private sector, even within the ranks of companies with knowledge of the upstream and provider-sourced data collection efforts of the National Security Agency.

With AI, reliance on distributed smart infrastructure critical to both national security efforts and targets of foreign cyber-enabled manipulation exacerbates the traditional civil-military relations problem already in existence in the digital age. How does the government carry out its security mission and ensure its coercive capability when it is forced to cede ownership of that mission to the de facto governors—including technology companies, Internet service providers, and backbone operators—of the operational domain in question (cyberspace)? Naturally, this problem strikes at the heart of challenges encountered and problematized in recent years regarding attempts to deter foreign digital aggression via cost imposition by denial. The current strategy is, in many ways, a military-oriented solution to challenges that are not—as so many scholars and strategists are wont to suggest—purely driven by domain characteristics but also by legal, normative, and practical government-industry challenges to ensuring national security in democratic states. Persistence underlying more conventional deterrent, norm-building efforts essentially constitutes an effort to define the character of the battlespace, pushing American presence everywhere to shape adversary expectations. With AI, the promise and problem of poisoning the battlespace suggests a (potentially massive) wrinkle for broader American efforts to head a liberal world order, as systematic efforts aimed at subverting algorithmic processes across global society to serve US security objectives spark inevitable outrage. Beyond the obvious broader issues that such outrage might bring about for American foreign policy efforts, the implication is yet another tangled web set to complicate PE as the bedrock of cyber strategy. After all, without additional communications methods baked into the strategy beyond conflict actions themselves, how can democratic states—and particularly the United States—maintain stable deterrent conditions when high political considerations force decision-makers to limit assertive digital activities? Permanent engagement may seem theoretically necessary, but it seems unlikely to be perpetually possible where exogenous changes in political conditions or in the nature of the battlespace threaten. AI stands to produce both.

The Learning Problem

Cyber conflict driven by the adaptability and rapidity brought on by AI poses several challenges to the strategy of persistent engagement. Policy makers 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. Moreover, state interest in operations of a poisoning nature via cyberspace is likely to grow over time as opportunities for manipulating processes that underlie strategy development and force posture determination proliferate.[55] 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 equal, the existence of robust AI systems on the part of foreign adversaries implies a learning problem: the more security institutions operate to shape behavior, the more adversaries should be empowered to understand and overcome such strategies. Much as in the case of generative adversarial networks (GAN) that study the actions AI models take to continually improve offensive capabilities, AI-enabled cyber forces presented with unique patterns of behavior-shaping attack from abroad will naturally undergo a process of adversarial learning.[56] Foreign action does not so much bound the shape of acceptable behavior as define the criteria under which future aggression is probabilistically less likely to induce some cost. Given the incentive described above toward the use of AI-enabled software agents with dramatically higher track records of success 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 PE such that defense practitioners might determine its effectiveness.[57] Such challenges multiply given the AI-ification of cyber conflict processes and the problem of poison as regular features of opera-

tions 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 significant fashion in an era where counteraction from foreign peers is not expected to be tit for tat, but rather an entirely alternative approach. In other words, where the paradigm of operations 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.

AI and Cyber Conflict Cascades

A final consideration seems particularly worthy of mention at this juncture. As is true in all areas of human interaction, misperception in cyber conflict is naturally not always—or even usually—a one-off occurrence. One action produces an interpretation of that action, which then informs further activity (or is itself that further activity). That reaction is then interpreted in turn, and so on. Misperception can spiral from minor assumption to major failure of interpretation if such a chain of events cannot be stopped. Such failures characterize many of the major conflict episodes in modern history. Of course, in strategic competition between states, one generally assumes that a great many analytic and procedural mechanisms bound up in the complex institutional landscape of international relations serve to backstop spiraling misperception.

Scholars have paid the problem of conflict spirals in cyberspace some sizable amount of attention, not least in the ubiquitous recognition that intention is difficult to ascertain in digital interactions. What may appear to be an attack may simply have been a probe, an effort to understand the battlespace or to engage in a non-warfighting activity. Beyond this level of discussion, however, scholars have given limited attention to the idea of cascading effects. After all, though automated attacks present a particular challenge wherein automated responses may be triggered, cascade effects at some point do tend to cease due to backstops in the algorithm—kill switches or conditional code that end a process without further human interaction.

With the use of AI, there is substantial risk that more interactions might produce a critical mass of activity leading to major unintended effects. One commonly cited example of such a critical mass event is the flash crash of the stock market on 6 May 2010. Though no definitive cause has been agreed upon by researchers, conventional wisdom attributes a Dow Jones

loss of almost 1,000 points in just 36 minutes to automated selling algorithms that reacted to an unusual perturbation of the market—often said to be an accidental sale some orders of magnitude above what was deemed normal. The result was a trillion-dollar loss in the market that then quickly rebounded in the following hours. Looking at the event, it is easy to imagine how dueling AI—or, perhaps more worryingly, a "battle" between AI and dumber automated algorithms—could rapidly and disastrously produce negative effects of strategic consequence. These could range from critical infrastructure shutdowns to counteroffensive cyber volleys of sufficient scale to prompt a state response beyond the domain.

Though this article does not attempt to address the challenge of cascades specifically, it seems clear that planners should avoid formalizing PE-style strategies in procedure and in code. Doing so would invite the opportunity for a diverse prospective set of flash crashes. It also seems reasonable to suggest that national security planners must be mindful of opportunities for such spiraling beyond the practice of cyber conflict. If CAIAs are indeed likely to become the norm of engagement in cyberspace, then we must be consistently mindful of the possibility that unexplained conflict developments not thought to be linked to the fifth domain may yet be affected by it. Thus, the human in the loop must not only be a decision-maker at US Cyber Command, but rather must also represent an assemblage of those stakeholders with jurisdiction over other areas of national defense.

Conclusion

The purpose of this article has been to contribute to the nascent literature on AI and national security activities by outlining how AI is likely to alter the shape and strategic calculations involved in interstate cyber conflict. It is hoped this information will be a resource for those interested in thinking more clearly about how AI stands to alter the dynamics of interstate and cyber conflict processes. Naturally, a substantial part of the effort here has been definitional. Indeed, it is from this effort (i.e., the categorization of different threat forms linked to the augmentation of cyber conflict processes by AI models and systems) that the primary argument of this article emerges.

Broadly, that argument is that the centrality of cyberspace to the deployment and operation of soon-to-be-commonplace AI systems implies new motivations for operations within the domain. More specifically, though AI does not itself imply inevitable advantages for attackers over defenders (or vice versa), adversarial learning techniques add complexity to already complex operational conditions in cyberspace and may contrib-

ute to an uptick in offensive behavior. Perhaps more worryingly, the centrality of the Internet to new AI systems incorporated across all areas of national security—not just to cyber conflict processes—indicates that sophisticated adversaries may be increasingly motivated to launch offensive online actions to achieve effects in other domains. The implications for current cyber conflict strategies—particularly those by Western defense enterprises—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 policy makers 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 opportunity 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 opportunities for the poisoning of 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 policy makers and military practitioners really know what to believe about strategic conditions?

Second, success in meeting the challenges of deploying AI for national security purposes will likely hinge on the approach organizations take toward trusting their AI systems and managing the interaction of human and machine operators.[58] To some degree the previous discussion involves the problem of "ghosts in the machine." That is, human assumptions present in the code of machine intelligence systems are the true problem underlying effective AI deployment for national security purposes. While such problems are arguably unavoidable as we move toward more common employment of AI, it seems likely that protocols for keeping humans in the loop at critical junctures are part of the solution to problems of system poisoning (either malicious or self-afflicted).

Finally—and perhaps most significantly—in the forthcoming era of AI-enabled contestation in world affairs, it seems clear that strategy development, assessment, and validation must emerge from the cross-domain understanding of the strategic motivations of adversaries. Cyberspace is not only a domain where unique forms of contestation and signaling can occur but also potentially the most critical terrain over which actions can be taken to affect processes that underlie all areas of modern society. Given

this potential, strategic planners would do well to build from assumptions that move beyond simple logic-of-the-domain characterizations of digital affairs. As some scholars have increasingly argued in both implicit and explicit terms, cyber conflict so often manifests in aid of nondigital contestation that we would do well to couch our analyses in terms of the logic of conflict processes other than cyber.[59] This stands to be especially the case with AI, not least given the fact that its targeting for security purposes is so likely to be tied to the use of computer and Internet systems upon which such programming must inevitably run. SSQ

Christopher Whyte, PhD

Dr. Whyte is an assistant professor at the L. Douglas Wilder School of Government and Public Affairs, Virginia Commonwealth University. His research interests include cyber conflict, information warfare, and emerging technology. He was lead editor for *Information Warfare in the Age of Cyber Conflict* (Routledge, 2020) and is co-author of a forthcoming Georgetown University Press book on military innovation surrounding artificial intelligence. An earlier version of this article appeared in the *Proceedings of the 12th International Conference on Cyber Conflict: 20/20 Vision: The Next Decade*.

Notes

1. It should be noted that the topic of involving AI 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 twentieth century.

2. See, among others, Stephen Hawking et al., "Transcendence Looks at the Implications of Artificial Intelligence—but Are We Taking AI Seriously Enough?," *The Independent*, 1 May 2014, https://www.independent.co.uk/; and Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (New York: Knopf, 2017).

3. For example, 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—in an attempt to convince audiences of the stakes involved in getting AI "right." 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, https://www.nickbostrom.com/.

4. The "Third Offset" is a strategy intended to be used by the US Department of Defense 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, deputy secretary of defense (speech, "Third Offset Strategy," Brussels, Belgium, 28 April 2016), https://www.defense.gov/; Cheryl Pellerin, "Deputy Secretary: Third Offset Strategy Bolsters America's Military Deterrence," *Defense News*, 31 October 2018, https://www.defense.gov/; and Katie Lange, "3rd Offset Strategy 101: What It Is, What the Tech Focuses Are," *DODLive* (blog), Defense Department, 30 March 2016, https://www.doncio.navy.mil/.

5. 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 (Stanford, CA: Stanford University Press, 2010); and Benjamin Jensen, "The Role of Ideas in Defense Planning: Revisiting the Revolution in Military Affairs," *Defence Studies* 18, no. 3 (2018): 302–17, https://doi.org/10.1080/14702436.2018.1497928. 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).

6. For the limited work to date on AI and strategic studies, see, for example, Benjamin M. Jensen, Christopher Whyte, and Scott Cuomo, "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence," *International Studies Review*, 2019, viz025, https://doi.org/10.1093/isr/viz025; 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)* (Tallinn: NATO CCD COE Publications, 2019), 249–65, https://ccdcoe.org/; and Kareem Ayoub and Kenneth Payne, "Strategy in the Age of Artificial Intelligence," *Journal of Strategic Studies* 39, no. 5-6 (2016): 793–819, https://doi.org/10.1080/01402390.2015.1088838; Heather Roff, *Advancing Human Security through Artificial Intelligence* (London: Chatham House, May 2017), https://www.chathamhouse.org/; Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review* 1, no. 3 (May 2018): 36–57, https://doi.org/10.15781/T2639KP49; Kenneth Payne, *Strategy, Evolution, and War: From Apes to Artificial Intelligence* (Washington, D.C.: Georgetown University Press, 2018); Heather M. Roff, "COMPASS: A New AI-Driven Situational Awareness Tool for the Pentagon?," *Bulletin of the Atomic Scientists*, 10 May 2018, https://thebulletin.org/; Kenneth Payne, "Artificial Intelligence: A Revolution in Strategic Affairs?," *Survival* 60, no. 5 (2018): 7–32, https://doi.org/10.1080/00396338.2018.1518374; Michael Horowitz et al., "Strategic Competition in an Era of Artificial Intelligence," Center for a New American Security (CNAS), 25 July 2018, https://www.cnas.org/; and Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," February 2018, *arXiv:1802.07228,* https://img1.wsimg.com/.

7. See, for instance, Enn Tyugu, "Artificial Intelligence in Cyber Defense," in *Proceedings of the 2011 3rd International Conference on Cyber Conflict*, eds. C. Czosseck, E. Tyugu, and T. Wingfield (Tallinn: NATO Cooperative Cyber Defence Centre of Excellence, 2011), 95–105; and Mariarosaria Taddeo and Luciano Floridi, "Regulate Artificial Intelligence to Avert Cyber Arms Race," *Nature* 556 (April 2018): 296, https://media.nature.com/.

8. For a broad overview of the scope and dynamics of cyber conflict, see, for example, Brandon Valeriano and Ryan C. Maness, *Cyber War versus Cyber Realities: Cyber Conflict in the International System* (Oxford: Oxford University Press, USA, 2015); and Christopher Whyte and Brian Mazanec, *Understanding Cyber Warfare: Politics, Policy and Strategy* (Abingdon: Routledge, 2018).

9. Machine learning is technically a subfield of AI research that, according to many, now virtually demands consideration as its own technology.

10. For instance, Brundage et al., "Malicious Use of Artificial Intelligence."

11. Jensen, Whyte, and Cuomo, "Algorithms at War," 10.

12. For a contemporary description of such efforts, see, for example, Alan Turning, "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, https://pdfs.semanticscholar.org/.

13. 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.

14. See Kline, 5–16; J. Moor, "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years," *AI Magazine* 27, no. 4 (Winter 2006): 87–91, https://www.aaai.org/; and Bruce Buchanan, "A (Very) Brief History of AI," *AI Magazine* 26, no. 4 (Winter 2005): 53–60, https://www.aaai.org/.

15. J. McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," 31 August 1955, http://www-formal.stanford.edu/.

16. Burton and Soare, "Understanding the Strategic Implications," 5.

17. Jensen, Whyte, and Cuomo, "Algorithms at War."

18. For an overview of machine learning, see Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning" *Nature* 521 (May 2015): 436–44. Also see Volodymyr Mnih et al., "Human-Level Control through Deep Reinforcement Learning," *Nature* 5, no. 18 (2015): 529–33, http://dx.doi.org/10.1038/nature14236; and David Silver et al., "Mastering the Game of Go without Human Knowledge," *Nature* 550, no. 7676 (October 2017): 354–59, https://doi.org/10.1038/nature24270.

19. 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* (New York: Basic Books, 2015).

20. For a seminal description of perception as a component element of broader attempts to build deep learning and reasoning systems, see Nicola Jones, "Computer Science: The Learning Machines," *Nature* 505, no. 7482 (2014): 146–48, https://www.nature.com/.

21. For further information on NLP, see Stephen F. DeAngelis, "The Growing Importance of Nature Language Processing," *Wired*, February 2014, https://www.wired.com/; 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, https://doi.org/10.1109/MCI.2014.2307227.

22. For further reading on intelligent machine vehicle systems, see Mario Gerla et al., "Internet of Vehicles: From Intelligent Grid to Autonomous Cars and Vehicular Clouds," *2014 IEEE World Forum on Internet of Things* (*WF-IoT)*, Seoul, 2014, 241–46, https://doi.org/10.1109/WF-IoT.2014.6803166; and Alberto Broggi et al., "Intelligent Vehicles," in *Springer Handbook of Robotics*, 2d ed., eds. Bruno Siciliano and Oussama Khatib (Berlin: Springer, 2016), 1627–56, https://link.springer.com/.

23. "Attack surface" is a term of art used to describe the sum of weak points of a given system. According to Tim Stevens, "the attack surface is less a physical boundary to be defended than a logical membrane of potential vulnerability distributed in space and time." More than just a set of functional components, an attack surface typically includes these elements: technical (i.e., infrastructure), social (i.e., the behaviors and psychology of system users/operators), and economic (i.e., the competing interests that characterize a system's usage). Tim Stevens, "Knowledge in the Grey Zone: AI and Cybersecurity," *Digital War*, 2020, https://www.researchgate.net/.

24. For a description of the episode in context, see DarkTrace, *The Next Paradigm Shift: Cyber-Attacks, AI-Driven*, research white paper (San Francisco: DarkTrace, 2018), https://www.oixio.ee/. 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 (January 2019): 17–19, https://doi.org/10.1016/S1361-3723(19)30010-7.

25. 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, https://doi.org/10.1080/09636412.2015.1038188.

26. 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* (Cambridge, MA: Belfer Center for Science and International Affairs, 2019), https://www.belfercenter.org/, 28.

27. Comiter, 19.

28. See Norman B. Hutcherson, *Command and Control Warfare: Putting Another Tool in the War-Fighter's Data Base*, No. AU-ARI-94-1 (Maxwell AFB, AL: Air University Press, 1994), https://apps.dtic.mil/; and Jeffrey A. Harley, *The Role of Information Warfare: Truth and Myths* (Newport, *RI:* Naval War College, Joint Military Operations Dept. 1996), https://apps.dtic.mil/.

29. See Comiter, *Attacking Artificial Intelligence*, 28.

30. See recent work, for instance, Ali W. Shafahi et al., "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks," presented at the 32nd Conference on Neural Information Processing Systems (NIPS), Montréal, Canada, 2018, https://arxiv.org/; Pang Wei Koh, Jacob Steinhardt, and Percy Liang, "Stronger Data Poisoning Attacks Break Data Sanitization Defenses," *arXiv preprint arXiv:1811.00741* (2018), https://arxiv.org/; Saeed Mahloujifar and Mohammad Mahmoody, "Can Adversarially Robust Learning Leverage Computational Hardness?," *arXiv preprint arXiv:1810.01407* (2018), https://arxiv.org/; and Chen Zhu et al., "Transferable Clean--Label Poisoning Attacks on Deep Neural Nets," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019, 7614–23, *arXiv preprint arXiv:1905.05897* (2019), http://proceedings.mlr.press/.

31. Shafahi et al., "Poison Frogs!"

32. Matthew Jagielski et al., "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," in *2018 IEEE Symposium on Security and Privacy (SP)* (New York: IEEE, 2018), 19–35, https://arxiv.org/.

33. Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz, "Prediction Poisoning: Utility-Constrained Defenses against Model Stealing Attacks," *arXiv preprint arXiv:1906.10908* (2019), https://arxiv.org/.

34. The White House, *National Cyber Strategy of the United States of America* (Washington, D.C.: The White House, 2018), https://www.whitehouse.gov/.

35. Paul M. Nakasone, "An Interview with Paul M. Nakasone," *Joint Force Quarterly* 92 (1st Quarter 2019): 4–9, https://ndupress.ndu.edu/.

36. Michael P. Fischerkeller and Richard J. Harknett, "Deterrence Is Not a Credible Strategy for Cyberspace," *Orbis* 61, no. 3 (2017): 381–93, https://doi.org/10.1016/j.orbis.2017.05.003.

37. For the broad literature on deterrence in cyberspace, see, for example, Martin C. Libicki, *Cyberdeterrence and Cyberwar* (Santa Monica, CA: RAND Corporation, 2009), https://www.rand.org/; Amir Lupovici, "Cyber Warfare and Deterrence: Trends and Challenges in Research," *Military and Strategic Affairs* 3, no. 3 (2011): 49–62, https://pdfs.semanticscholar.org/; Matthew D. Crosston, "World Gone Cyber MAD: How "Mutually Assured Debilitation" Is the Best Hope for Cyber Deterrence," *Strategic Studies Quarterly* 5, no. 1 (2011): 100–116, https://deepsec.net/; Eric Talbot Jensen, "Cyber Deterrence," *Emory Int'l L. Rev.* 26 (2012): 773, https://law.emory.edu/; Dorothy E. Denning, "Rethinking the Cyber Domain and Deterrence," *Joint Force Quarterly* 77 (2d Quarter 2015): 8–15, https://ndupress.ndu.edu/; Emilio Iasiello, "Is Cyber Deterrence an Illusory Course of Action?." *Journal of Strategic Security* 7, no. 1 (2014): 54–67, https://scholarcommons.usf.edu/; and Uri Tor, " 'Cumulative Deterrence' as a New Paradigm for Cyber Deterrence," *Journal of Strategic Studies* 40, no. 1-2 (2017): 92–117, https://doi.org/10.1080/01402390.2015.1115975.

38. Michael P. Fischerkeller and Richard J. Harknett, "Persistent Engagement, Agreed Competition, Cyberspace Interaction Dynamics and Escalation," *Orbis* 61, no. 3 (Summer 2017): 381–93.

39. See, for example, Department of Defense, Defense Science Board, *Defense Science Board Task Force on Cyber Deterrence* (Washington, D.C.: Defense Science Board, 2017), https://apps.dtic.mil/; and Amb. John Bolton, "Transcript: White House Press Briefing on National Cyber Strategy – Sept. 20, 2018," https://news.grabien.com/.

40. 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* (Arlington, VA: Cyber Conflict Studies Association, 2013).

41. See, for example, Nicholas Tsagourias, "Cyber Attacks, Self-Defence and the Problem of Attribution," *Journal of Conflict and Security Law* 17, no. 2 (2012): 229–44, https://papers.ssrn.com/; Jon R. Lindsay, "Tipping the Scales: the Attribution Problem and the Feasibility of Deterrence against Cyber attack," *Journal of Cybersecurity* 1, no. 1 (2015): 53–67, https://doi.org/10.1093

/cybsec/tyv003; and Thomas Rid and Ben Buchanan, "Attributing Cyber Attacks," *Journal of Strategic Studies* 38, no. 1-2 (2015): 4–37, https://ridt.co/.

42. 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* (New York: Oxford University Press, 2016).

43. See Jason Healey, "The Implications of Persistent (and Permanent) Engagement in Cyberspace," *Journal of Cybersecurity* 5, no. 1 (2019): tyz008, https://doi.org/10.1093/cybsec/tyz008.

44. Healey, 5.

45. Healey, 5–6. For the original discussion of the notion of persistence feeding deterrent norm-building, see Fischerkeller and Harknett, "Persistent Engagement," 388–91.

46. Among others, see Max Smeets, "Cyber Command's Strategy Risks Friction with Allies," *Lawfare*, blog, 28 May 2019, https://www.lawfareblog.com/; Brandon Valeriano and Benjamin Jensen, "The Myth of the Cyber Offense: The Case for Restraint," CATO Institute Policy Analysis 862, 15 January 2019, https://www.cato.org/; Herb Lin and Max Smeets, "What Is Absent from the U.S. Cyber Command 'Vision,'" *Lawfare*, blog, 3 May 2018, https://www.lawfareblog.com/; and Max Smeets and H. A. Lin, "A Strategic Assessment of the U.S. Cyber Command Vision," in *Bytes, Bombs and Spies: The Strategic Dimensions of Offensive Cyber Operations*, eds. Herbert Lin and Amy Zegart (Washington, D.C.: Brookings Institution Press, 2019), http://www.jstor.org/.

47. Healey, "Persistent (and Permanent) Engagement," 7–8.

48. Michael P. Fischekeller and Richard J. Harknett, "What Is Agreed Competition in Cyberspace?," *Lawfare*, blog, 19 February 2019, https://www.lawfareblog.com/.

49. Even the head of Cyber Command admits this. See "An Interview with Paul M. Nakasone," *Joint Force Quarterly* 92 (1st Quarter 2019): 4–9, https://ndupress.ndu.edu/.

50. As the wording of the strategy itself suggests, taking reference from Fischerkeller and Harknett's original analysis in "Deterrence Is Not a Credible Strategy for Cyberspace."

51. Max Smeets, "There Are Too Many Red Lines in Cyberspace," *Lawfare*, blog, 20 March 2019, https://www.lawfareblog.com/.

52. Valeriano and Jensen, "Myth of the Cyber Offense," 8. This argument is based on the original notion of the "fleet-in-being" developed by Corbett. See Julian S. Corbett, *Some Principles of Maritime Strategy*, ed. Eric Grove (Annapolis: US Naval Institute, 1988).

53. See the recent US Cyberspace Solarium Commission report for a description on prevailing thought on the relationship between cost imposition via persistent engagement, deterrent operations, and norm building. Angus King and Mike Gallagher, co-chairs, US Cyberspace Solarium Commission, *Cyberspace Solarium Commission Report* (Arlington, VA: US Cyberspace Solarium Commission, March 2020), 7, https://www.solarium.gov/.

54. Healey, "Implications of Persistent (and Permanent) Engagement," 25.

55. This assertion is quite arguably backed by work that demonstrates in both quantitative and qualitative terms in increasing turn toward 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* (New York: Oxford University Press, 2018).

56. For GANs, see James Vincent, "Deepfake Detection Algorithms Will Never Be Enough," *The Verge*, 27 June 2019, https://www.theverge.com/. The phrase "adversarial learning" is a common one used 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 Daniel Lowd and Christopher Meek, "Adversarial Learning," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 641–47, ACM, 2005; and Pavel Laskov and Richard Lippmann, "Machine Learning in Adversarial Environments," *Machine Learning* 81, no. 2 (2010): 115–19, https://doi.org/10.1007/s10994-010-5207-6.

57. 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, https://doi.org/10.23919/CYCON.2019.8756890.

58. 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, https://doi.org/10.1162/ISEC_a_00267.

59. 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, https://doi.org/10.1093/isr/viw013.