MAY 2020

A National Security Research Agenda for Cybersecurity and Artificial Intelligence

CSET Issue Brief

AUTHOR
Ben Buchanan

Introduction

The intersection between cybersecurity and artificial intelligence is ripe for serious study from a variety of angles. There are purely technical aspects of great importance, such as how artificial intelligence changes the discovery of software vulnerabilities useful for hacking computer systems and the capacity for defenders to detect malicious code within their networks. Yet many of these technical questions have already been well-specified and are the subject of promising inquiries. This research agenda instead examines a different angle, one of national security.[1]

A national security-driven research agenda is informed by technical evidence, but not limited by it. It considers how the balance of technical facts shapes questions likely to matter to national security policymakers and scholars who would otherwise overlook the technology. More generally, it offers policymakers a set of questions—and, someday, answers—that they should consider, but that are probably unfamiliar to them.

This agenda focuses on the machine learning paradigm of artificial intelligence. It has four components: offense, defense, adversarial learning, and overarching questions.

- Offense considers the ways in which machine learning might change the techniques adversaries already use to gain unauthorized access to computer systems, from discovering software vulnerabilities to infiltrating a target system and beyond.
- Defense considers how machine learning systems can aid in detecting and responding to intrusions, as well as remediating malicious code.
- Adversarial learning examines the cybersecurity weaknesses of machine learning systems themselves and the data upon which they depend.
- Overarching questions examine the ways in which the properties and powers of machine learning systems can change the strategy and

conduct of cyber operations, from influence campaigns to accident risks to strategic stability and beyond.

Each section contains several key questions, including some whose answers will be complex and context-dependent; the discussion that follows is an illustrative, rather than exhaustive, list.

Offense

Offensive cyber operations have long been conceptualized using an idea known as the kill chain—the sequence of steps hackers cycle through in order to achieve their aims. One of the foremost national security questions at the intersection of cybersecurity and AI is the degree to which machine learning will reshape or supercharge this kill chain. There are reasons for concern, but also reasons to think present-day automation—not using machine learning techniques—is already effective in human-machine teams; perhaps the additional power offered by machine learning may not tip the scales. A thorough study of the kill chain is in order.

Such a study may be of particular value to network defenders or software engineers. For example, both hackers and defenders have an interest in finding vulnerabilities in software; the former to exploit them, the latter to remediate them.

Vulnerability Discovery

Can machine learning better find software vulnerabilities? A fundamental aspect of many offensive cyber operations is finding and exploiting these weaknesses in computer code. While many of the vulnerabilities exploited in cyber operations have been found and used by others, some are novel. These so-called zero-day vulnerabilities confer greater power because no security patches exist to stop hackers from exploiting them. While zero-days are often overhyped in cybersecurity policy discussions, finding and exploiting them remains a key part of advanced modern cyber operations. All else being equal, hackers who are more capable of doing this—and of stringing zero days together into exploit chains achieving still greater access—will have more freedom of action and offensive capability.

Automated tools can already help find vulnerabilities that might be exploitable. In particular, tools known as fuzzers provide carefully crafted

inputs to computer code, seeking failures that would reveal a vulnerability. Researchers are exploring how machine learning might improve the analysis of data generated by fuzzers and find vulnerabilities that would go undiscovered using current methods.[2] In the hands of skilled hackers, better fuzzers and better tools for analyzing the data they produce will likely confer an advantage.

Spear-phishing

Can machine learning better tailor and scale spear-phishing attempts? Spear-phishing—the practice of delivering malicious code or gaining unauthorized access via socially engineered messages—remains one of the most common and effective offensive techniques. A panoply of notable cyber operations have relied on it, from the Russian hack of Clinton campaign chairman John Podesta's emails to reams of Chinese espionage efforts. And yet the process of spear-phishing can seem manual and cumbersome: finding a target, determining what kind of message the target might believe, then crafting and sending such a message. The less-careful alternative to spear-phishing, known simply as phishing, forgoes much of this customization to achieve greater scale; in phishing operations, many more messages are sent with much less sophistication, such as the once-ubiquitous claim of money waiting in a bank account in Nigeria.

AI might offer the possibility to retain the relative sophistication of spear-phishing while also attaining the scale of traditional phishing. If machine learning systems can generate credible messages that appear to come from plausible senders—and that evade any automated attempts at detection—they could materially increase the volume of such messages and their potential success rate. Achieving this would require substantial advances in natural language processing, an area of inquiry that has seen rapid growth over the past decade.[3]

Propagation

Can machine learning change how effectively malicious code spreads itself? Self-propagating cyber capabilities have been around for decades, dating to at least the 1988 Morris worm. The ability of malicious code to spread itself from computer to computer and network to network is worrying, as it offers exponential reach in offensive cyber operations. A number of attacks with national security implications have exploited this potential, including the

Stuxnet worm that targeted Iran's nuclear program and the 2017 NotPetya attack that affected hundreds of thousands of computers in more than 100 countries, causing more than $10 billion in damage.

NotPetya's automated propagation mechanism, which relied mostly on password theft but also on a repurposed National Security Agency (NSA) exploit, meant it could rip through targeted networks in seconds or minutes, making it one of the fastest-spreading pieces of malicious code in history. One analyst of the code, Craig Williams of Cisco, said, "by the second you saw it, your data center was already gone." NotPetya did not employ any machine learning techniques to achieve this prolific speed.[4]

Given how impressive baseline self-propagation mechanisms are without AI, it remains an open question whether machine learning will render future attacks any more effective. Current techniques usually rely on only a handful of propagation mechanisms and succeed primarily because so many systems are vulnerable. More complex mechanisms might be necessary for self-propagation of malicious code if a defender employs improved machine learning-enabled protections. One option for attackers could be to enable a greater number and diversity of propagation techniques based upon actual network conditions encountered during an intrusion. Machine learning mechanisms might also enable better selection of propagation techniques compared to present-day automation and heuristics. That said, both of these claims are speculative and need further investigation, ideally with quantifiable metrics and perhaps with lab experiments. It may be that, for all of their hype, machine learning systems will not meaningfully change the self-propagation of malicious code and that incentives for pursuing new and more advanced techniques will remain low if current ones continue to work well.

Obfuscation and Anti-Forensics

Can machine learning better hide offensive cyber operations? One of the primary goals of an intruder is not to get caught. If machine learning continues to aid defenders in detecting malicious activity—and it will likely improve at that—perhaps it also might aid hackers looking to remain undetected. Intruders have long taken steps, from code obfuscation to packing to process hollowing to fileless malicious code and more, to try to reduce the visibility of their operations. In theory, machine learning systems could aid the deployment of these anti-forensics tools. At a minimum, it will be necessary for intruders to understand the weaknesses of the machine

learning-based defenses they will face and to exploit those weaknesses, perhaps using adversarial learning as discussed below.

Destructive Power

Can machine learning make cyber capabilities more powerful? In general, the most destructive cyber capabilities are the ones that take existing components of physical and computer systems and either prevent them from functioning or cause them to work in unintended ways.[5] For example, wiping attacks cripple computing infrastructure by overwriting fundamental pieces of computer code on which the system rests. Stuxnet and other attacks against critical infrastructure manipulated commands in order to destroy physical components, such as by changing the speed at which a centrifuge spins or the pressure of the gas within it.

Machine learning is unlikely to offer much in wiping attacks. If hackers obtain sufficient privileges and access, executing a wiping attack is not technically sophisticated and leaves little room for innovation. But, manipulating complex physical infrastructure, seems more promising. The widespread use of systems modeling employed by high-end manufacturers to monitor and maintain their products could provide fertile ground for AI systems to facilitate physically destructive attacks. By repurposing a stolen systems model or understanding a system's configuration, machine learning could help calibrate and camouflage physically destructive attacks in order to slowly render systems unusable. It is here that machine learning capabilities might make a substantial difference in the power of offensive operations, though the data remains extremely sparse and the concept speculative.

Perhaps the most salient piece of data is the 2016 blackout in Ukraine caused by Russian hackers working for the GRU, Russia's military intelligence agency. Unlike the 2015 blackout in Ukraine, which was manually orchestrated in a step-by-step fashion, the 2016 blackout employed automation, though not machine learning. The code, known as CRASHOVERRIDE, contained components that sought to identify key aspects of the targeted power systems and then to enable an attack.[6] Though not all components functioned as intended, it is perhaps a harbinger of things to come, in which automated attack systems yield cyber operations more powerful in their destructive physical effects.

Defense

Machine learning holds promise for cyber defense. The single biggest challenge for network defenders is detection: finding the adversary's presence in one's own network. Detection times vary based on the sophistication of the attacker and defender, but the average once lingered at well over a year. While defenders have improved, in many cases intruders can operate for months within the target network, unnoticed and unconstrained.[7] Virtually every major cyber attack—such as Stuxnet, the two blackouts in Ukraine, and NotPetya—has been preceded by months, if not years, of reconnaissance and preparation.[8] This window offers an opportunity. If machine learning can improve detection, interdiction, and attribution, it can dramatically reduce the potential dangers of cyber operations. That said, machine learning has been applied to cyber defense for several years already and challenges persist; it is thus vital to ground the evaluation of machine learning-aided cyber defense not just in theory but in practical—and ideally measurable—results.

It is worth noting again that some offensive technology has defensive applications as well, insofar as defenders choose to simulate offensive actors or find and remediate software vulnerabilities before hackers discover and exploit them.

Detection

Can machine learning help detect malicious code when it arrives on a network or a computer system? Detection is the first fundamental challenge for cyber defense; if an adversary is not found, it cannot be removed, and many cyber operations have gone unimpeded for months or years because of a failure to detect.

Machine learning may substantially improve detection. The amount of data on modern computer systems and networks is so vast that most non-AI techniques, including human analysis, cannot keep up. While humans must still investigate and adjudicate more complex alerts, machine learning systems can make continual first passes through data in order to detect anomalous activity. In addition, machine learning techniques such as supervised learning have proven capable of finding patterns in large, complex data sets that have previously remained hidden from human analysis and traditional techniques. Engagement with private sector analysts will likely shed substantial light on

the usefulness of this technology and the ways in which it may improve. Unlike many of the other possible applications of machine learning discussed in this research agenda, substantial data supports this claim, though that data is often proprietary in its nature.

Interdiction

Can machine learning help thwart, not just detect, offensive cyber operations? It stands to reason that if machine learning can detect offensive cyber operations in progress, then it can probably help delay or block them as well. This assumption requires a high confidence that the machine learning system can function as intended, since the effects of an inadvertent shutdown can be substantial. Nonetheless, if that confidence is warranted, then it is quite possible to extend detection capabilities to include defense, stopping or otherwise interfering with malicious actors as they begin interacting with target systems.

The best example of this hope may be DARPA's Grand Cyber Challenge in 2016, which pitted automated competitors against one another in a stylized capture the flag competition. In addition to gaining access to other competitors' systems, the players had to defend their own systems, finding and addressing weaknesses in code. That said, the DARPA challenge, though significant, is hardly a real-world example, and public investigation of this topic has waned since then. Much more analysis is needed about the concrete capabilities, risks, and tradeoffs involved in using machine learning for more proactive cyber defense.

Attribution

Could machine learning programs more effectively attribute cyberattacks? While fundamentally a political matter, attribution has substantial technical inputs. It is also one of the most debated—and perhaps most important—subjects in cybersecurity strategy and a foundation for goals like deterrence.[9] While attribution is often more tractable than many believe, machine learning might strengthen it further.

For example, unsupervised learning clustering algorithms could link code snippets, identify persistent groups of attackers, or help identify groups that could be responsible for a new attack. Or natural language processing might allow for linguistic attribution of code snippets or of messages; such methods,

for example, might have determined that the specific grammatical errors made in the WannaCry English messages most probably came from native Korean speakers. Advanced cybersecurity companies likely already employ some of these methods in their analysis.

Adversarial Learning

Thus far, this research agenda has focused on how machine learning will change the current and future practice of offensive and defensive cyber operations. But another set of questions also deserves analysis: what about the cybersecurity vulnerabilities of machine learning systems themselves? It stands to reason that they will be vulnerable to many of the same weaknesses as traditional computer systems, such as the potential for software bugs that an attacker could exploit. Moreover, they offer new kinds of fundamental vulnerabilities providing hackers additional opportunities to undermine the effectiveness of machine learning systems in critical moments. Yet, for all of this, credible estimates suggest only one percent of AI research money is spent on machine learning security and safety.[10]

Adversarial Examples

Can machine learning systems be fooled, and can they be secured against these attempts at deception? One of the fundamental tasks of modern machine learning is classification: identifying to which category something fits. The field is full of these sorts of systems, from spam filters to digit readers to image recognition tools. Any good classification system needs to be alert to forgeries, camouflage, and deception. Some types of deception are quite intuitive; for example, a four may be written to look more like a nine, or a spam email might try to sound authoritative and credible. These deceptions mostly seek to fool humans by taking advantage of how the brain processes information.

But other deceptions aim to fool machine learning systems instead of, or in addition to, humans. Adversarial examples fit into this category. They target the mechanisms neural networks use to process information and make classifications, which are distinct from human information processing. Once understood—often by running the neural network in reverse—these mechanisms can be exploited. Hackers can craft an adversarial example that looks entirely normal to a human but like something very different to a machine. Examples from academic study show how changing just a few

pixels can affect how a machine learning system classifies an image, dramatically changing the result in a way almost entirely undetectable to human observers.[11]

As machine learning systems are used in more prominent and important decisions, two things occur. First, the number of potential hackers increases, as more people will have an incentive to target those systems. Second, the consequences of failure increase, as the decisions are by definition higher stakes. Machine learning systems may be used in intelligence analysis or even lethal autonomous weapons—contexts where managing the risks of adversarial examples becomes fundamental. The possibility of adversarial examples must be carefully studied before machine learning systems are deployed to any mission-critical environment and robust countermeasures must be developed.

Data Poisoning

Can machine learning systems fail due to their training data? For many machine learning systems, training data is fundamental. Without the explicit instructions provided in other forms of computer programs, these systems learn virtually all they will ever know from the data they are provided. For hackers, this creates an opportunity: change what data the system sees during training and therefore change how it behaves. This risk is particularly acute for systems continually trained and retrained based on user input, such as recommendation systems and some spam filters.

This class of activity, known as data poisoning, deserves further technical study.[12] More in-depth research will likely reveal significant vulnerabilities for machine learning systems deployed in national security environments in which there are dedicated adversaries.

Data Pipeline Manipulation

While data poisoning focuses on contaminating the training process for neural network classifiers, other malicious data attacks are relevant as well. Even without attacking the model and training data, an attacker can cause data to be misclassified by modifying the input data before it reaches the machine learning system. Poor underlying cybersecurity practices could allow attackers to modify input data while in transit or while stored on servers. Researchers demonstrated this attack vector recently by modifying CT scans

while the data was in transit from the CT scanning machine to the data server.[13]

It may be possible to gain access using regular cybersecurity vulnerabilities and then insert carefully generated or manipulated malicious data. Drawing on generative adversarial networks (GANs) or other deep fake technology, this data could fool both machine and human observers. GAN-created data inputs coupled with traditional cyber data integrity attacks can introduce misleading data into the processing stream and dupe both humans and supporting AI systems.

Model Inversion

Can machine learning systems unintentionally reveal secrets? Consider this possibility: a machine learning system is training on classified data, perhaps for an intelligence analysis task. It is then deployed to a real-world, unclassified environment in which it performs this task and analyzes adversary activity. By interacting with the machine learning model, subtly changing its activity, and using a technique known as model inversion, the adversary may be able to deduce key features of the underlying data on which the system was trained, essentially gaining access to classified secrets.

Model inversion remains mostly an academic topic according to open sources, but it merits further study.[14] The risk of machine learning systems unwittingly revealing secret information is significant enough to require mitigation before systems are deployed to environments with technically capable adversaries.

Overarching Questions

Policymakers should consider the degree to which machine learning systems in the cybersecurity domain will present overarching questions. These challenges will arise from the application of machine learning but are likely to be missed by research that is strictly technical in focus. At least five are immediately apparent and more will almost certainly follow.

Cyber Accidents

How will machine learning change the risk of accidents in cyber operations? The study of "normal accidents" is important in any complex field.[15] While it

garners a great deal of attention in nuclear strategy, it has received almost none in cyber operations. This oversight is surprising, given the good reasons to think cyber accidents will occur with regularity. Computer bugs and software failures are common enough even for legitimate operators of systems and without interference; unexpected operational failures will only increase when attackers try to write code that manipulates a system they do not understand well.

Notable cyber accidents have likely already occurred. For example, it seems probable that the 2017 attack known as WannaCry, carried out by North Korean hackers and causing more than $4 billion in damage all over the world, was at least partially unintentional. Though designed to act as ransomware, the attack contained no mechanism for decrypting files once a ransom was paid. The North Koreans may not have understood the power of an NSA exploit, ETERNALBLUE, which they repurposed and added to the code.[16]

There are other cases, too. A mysterious attack on a steel plant in Germany in 2014 was apparently an espionage operation gone wrong. It triggered automated shutdown procedures that damaged the facility.[17] The countrywide outage of Syria's internet in 2012 was purportedly another espionage attempt, this time by the NSA, that inadvertently disabled key internet routing functions.[18] Stuxnet, the reported U.S. and Israeli attack on Iran, propagated far further than intended, eventually turning a covert program into one of the most famous cyber operations ever.[19] More generally, the British signals intelligence agency GCHQ published internally a guide entitled, "What's the Worst That Could Happen?" that has since leaked, suggesting that other types of accidents could and likely have occurred, hidden only by classification.[20]

The advent of machine learning in cyber operations seems poised to make the risk of cyber accidents worse, not better. While human operators are certainly fallible, machine learning systems have particular kinds of failure modes that increase the accident risk. Indeed, many past accidents have involved an automated component, though not one that used machine learning. This reality could well be a harbinger of things to come.

Moreover, at least in the near term, machine learning capabilities will add complexity to traditional attack vectors, raising the risks that cyber operators may adopt machine learning features without fully understanding their inner

workings or potential effects. For example, machine learning systems could increase the risk of targeted computer systems causing errant shutdowns, as appeared to occur with the German steel mill. Or more automated offensive systems may scale up the cost of failures, spreading destructive malicious code further than intended, sometimes with substantial consequences, as both Stuxnet and WannaCry suggest. Guarding against these accidents—and spotting them when they do occur—will be essential.

Influence Campaigns

Can propaganda and influence be automated? There is substantial discussion of the role Russian Twitter and Facebook bots played in interfering with the 2016 election in the United States.[21] This terminology obscures the fact that people, not code, carried out most of the significant Russian activities; however, that might not be true for future operations. If machine learning systems can generate realistic and convincing deception campaigns with minimal effort, then it stands to reason that they will quickly become an arrow in the propagandist's quiver.

There is good reason to think that such automated deception will soon become possible. In 2019, the leading research lab OpenAI announced the creation of GPT-2, a tool for generating streams of credible text from any given input.[22] Upon the release of the tool—which OpenAI delayed because of national security concerns related to its underlying power—users from all over the world started demonstrating what it could do. They inputted initial sentence-length prompts and watched it quickly generate paragraphs of mostly credible text. A subsequent study published in *Foreign Affairs* demonstrated the salience for geopolitics; 72 percent of users thought "news" stories generated by GPT-2 were credible.[23]

These tools will require more steering and shaping in order to be useful in disinformation campaigns, but further technological development of natural language processing systems powered by machine learning seems likely. After unveiling GPT-2, OpenAI's subsequent work showed how the tools could respond to human direction in generating text.[24] More generally, there is ample appetite for what researchers call computational propaganda: the use of machines to amplify, scale, and shape information campaigns.

Speed

How will machine learning change the speed of cyber operations? Speed has long been a central focus for policymakers. Former White House official Richard Clarke claimed that "cyberwar happens at the speed of light," while former national security official Joel Brenner contended that, "speed, not secrecy, is the coin of the realm." Former Director of the NSA Keith Alexander told Congress that "in terms of...cyberattacks, it is over before you know what happened. These happen at lightning speed." Martin Dempsey, then-Chairman of the Joint Chiefs of Staff, made the comparison to humans explicit when he said that the military must be "able to operate at network speed, rather than what I call swivel-chair speed."[25]

Each of these individuals, and much of the conventional wisdom to this point, likely overstates the speed of cyber operations. Most operations remain human-directed and human-conducted, proceeding very much at the swivel-chair speed sometimes derided by senior officials. Thus far, the process of finding and exploiting software vulnerabilities, writing malicious code, selecting a target, gaining access to the target via spear-phishing or other means, establishing command and control, and moving through the network is in most cases a human process, though tools certainly help. Other components of cyber operations, such as legal reviews and bureaucratic authorizations, are perhaps even slower; they take place at committee speed.

But machine learning may change this dynamic. If, as outlined above, machine learning can automate key components of the kill chain, cyber operations could proceed much more quickly. If some forms of authority to take certain actions can be effectively delegated to the attack code, then operations might proceed with less cumbersome oversight and more quickly still. In short, machine learning may help enable the operational tempo that some policymakers had long imagined.

Offense-Defense Balance

Will machine learning benefit network intruders more than network defenders, or vice versa? The question of the offense-defense balance has both theoretical and practical relevance. In international relations theory, it is at the core of strategic stability, such as in the security dilemma discussed below; some argue that the offense-defense balance shapes nations' decisions to initiate conflict.[26] In cybersecurity practice, the offense-defense balance is

fundamental to the day-to-day business of securing computer networks. It shapes which techniques are likely to work on the offensive side and what level of capability and effort is required to carry out offensive cyber operations.

While theorists are fond of speaking of the offense-defense balance as a single variable, it is more accurate to think of it as dyadic, relevant to one specific pair of attacker and defender. These dyads can exist at the state level—the United States and China, for example—but also at the organizational one. Fundamentally, the ability to reap the offensive and defensive benefits offered by machine learning will depend on one's capacity to integrate the technology into already-existing procedures, data pipelines, and organizational capabilities. Organizations of a specific type, such as Wall Street banks spending half a billion dollars per year each on cybersecurity, may benefit from defensive cybersecurity advances, while less well-resourced organizations may not. The same is true for intruders; sophisticated intelligence agencies may be able to craft intricate automated tools while others fail to do so. Or the technology may diffuse broadly, rendering many of these organizational differences less important and leveling the playing field.

In addition, the offense-defense balance might vary depending on the objectives of the operation. It may be that some more basic kinds of operations are offense-dominant, but that others are defense-dominant; for example, machine learning might improve the ability to get access to computer systems, but also to detect intrusions, enabling faster efforts but making slower-burning campaigns against critical infrastructure much harder. In general, treating the offense-defense balance as a single variable in the context of machine learning in cybersecurity does not make sense. Much more context-specific research is needed to flesh out its complexities.

Proliferation

Are machine learning-enabled cyber capabilities more likely to be leaked, lost, or proliferated than previous types of weapons? On one hand, they might be more akin to nuclear weapons—espionage has occurred, but has never successfully delivered a viable, functioning weapon to a foreign state. Or, more likely, they might be more akin to cyber capabilities, which criminal groups often use after government hackers discover and deploy them.[27] They might be so portable as to enable the easy movement of many tools, the way

that an unknown group known as the Shadow Brokers proliferated a large quantity of NSA secrets.[28] Understanding the security and portability of machine learning-enabled cyber capabilities is essential for preventing their misuse and potentially for generating policy options to shape the proliferation or non-proliferation of capabilities.

Related to this central question of proliferation are questions about the role of medium-, small-, and non-state actors in the age of AI. The benefits of machine learning-enabled cyber capabilities may accrue to nations that are the most technologically sophisticated, in the way that only those sophisticated nations currently conduct attacks on industrial control systems. However, the capabilities may also be used by many states, in the way that commodity cyber espionage tools are widely circulated.

Strategic Stability

Will machine learning shape strategic stability in cyber operations? This is perhaps the most important emergent and cross-cutting concern. In effect, it is the aggregate of the component pieces of this research agenda, plus others that are as-yet unknown. Each of the factors outlined above will likely have strategic effects, yet what those are and how they interact with one another remains uncertain.

Theoretical tools can help guide research into strategic stability, but they are, as currently constituted, insufficient. For example, the security dilemma is the notion that as one state secures itself it unintentionally threatens other states, causing them to take steps to secure themselves and unintentionally threatening others. In some form, the security dilemma goes back to the ancient Greeks, when the historian Thucydides wrote, "It was the rise of Athens and the fear that this inspired in Sparta that made the Peloponnesian war inevitable."[29] Over the millennia since, the security dilemma has been formalized and developed, though many of its more-established components do not apply well to cyber operations.[30] It will likely need still further revision to be a useful mechanism for understanding strategic stability in the age of AI.

Other components of the research agenda will shape strategic stability as well. The importance of speed could increase the need for quick decisions based upon complex and incomplete information, perhaps raising the risk of misinterpretation; increased accident risks will make this possibility even more dangerous and render interpretation harder still. The growing power and

effects of increasingly automated cyber operations might raise the stakes, elevating these operations to levels of greater strategic concern. If offense-defense theory is to be believed, a reshaped balance wrought by machine learning will inform the strategic options favored by policymakers, perhaps leading to escalation or brinksmanship. The cybersecurity weaknesses of machine learning systems themselves, such as their susceptibility to adversarial examples, might further enhance their perceived frailty and contribute to a sense that policymakers must use them at the dawn of a conflict or risk losing them; such concerns about survivability of key systems do not augur well for stability.

On the other hand, machine learning in cyber operations might contribute to strategic stability, particularly if it can help defenders more than intruders. If broadly enjoyed by all, these defensive capabilities might shift the strategic environment so that nations are more secure in their cyber capabilities without needing to hack others. The security dilemma suggests that, though unlikely (and complicated by the fact that power differentials are likely to be dyadic, not global), this scenario would be very stable. More generally, machine learning might help to address some of the fundamental problems in cybersecurity, from vulnerability discovery during the development process to detection of malicious activity once underway. If it did so, it would help provide a stability not just geopolitical in nature, but one that extended to the technical ecosystem writ large.

It is important to appreciate the differences between nations in how they approach these questions. The American perception of the issues in this document, but especially those relating to geopolitical stability, likely diverges from the Chinese or Russian perception. Policy-relevant scholarship must bridge that gap by spotting areas of difference and potential misinterpretation. The national security research agenda for cybersecurity and AI is too important to view from just one perspective.

Acknowledgments

The author would like to thank John Bansemer, Dakota Cary, Teddy Collins, Wyatt Hoffman, Drew Lohn, Jason Matheny, Igor Mikolic-Torreira, Micah Musser, Chris Rohlf, Lynne Weil, and Alexandra Vreeman for their comments on an earlier draft of this agenda.

Endnotes

[1] Other research questions have been articulated in areas beyond national security. See for example Ben Buchanan and Taylor Miller, *Machine Learning for Policymakers: What It Is and Why It Matters* (Belfer Center for Science and International Affairs, 2017); Arjun Panesar, *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes* (Apress, 2019).

[2] Gary J. Saavedra et al., "A Review of Machine Learning Applications in Fuzzing," *arXiv [cs.CR]* (June 13, 2019), arXiv, http://arxiv.org/abs/1906.11133.

[3] Alec Radford et al., "Language Models Are Unsupervised Multitask Learners," *OpenAI Blog* 1, no. 8 (2019), https://openai.com/blog/better-language-models/.

[4] For this quote and broader discussion of NotPetya's speed, see Ben Buchanan, *The Hacker and the State: Cyber Attacks and the New Normal of Geopolitics* (Harvard University Press, 2020), chap. 13.

[5] Thomas Rid and Peter McBurney, "Cyber-Weapons," *RUSI Journal* 157, no. 1 (2012): 6–13.

[6] "CRASHOVERRIDE: Analysis of the Threat to Electric Grid Operations" (Dragos, June 13, 2017), https://dragos.com/blog/crashoverride/CrashOverride-01.pdf.

[7] "M-Trends 2020," FireEye, 11, https://content.fireeye.com/m-trends/rpt-m-trends-2020.

[8] Ben Buchanan, *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations* (Oxford University Press, 2017), chap. 2.

[9] Herbert Lin, "Attribution of Malicious Cyber Incidents: From Soup to Nuts," *Journal of International Affairs* 70, no. 1 (2016): 75–137; Thomas Rid and Ben Buchanan, "Attributing Cyber Attacks," *Journal of Strategic Studies* 39, no. 1 (2015): 4–37.

[10] This estimate is from Jason Matheny, former director of the Intelligence Advanced Research Project Activity and founding executive director of CSET. For more on the weaknesses of machine learning systems and a comprehensive taxonomy of attacks against AI architectures, see these two papers from the Berryville Institute of Machine Learning: Gary McGraw, Harold Figueroa, Victor Shepardson, Richie Bonett. 2020. "An Architectural Risk Analysis of Machine Learning Systems: Toward More Secure Machine Learning." Version 1.0. Berryville Institute of Machine Learning. Victor Shepardson, Gary McGraw, Harold Figueroa, Richie Bonett. "A Taxonomy of ML Attacks." Berryville Institute of Machine Learning. May 2019. https://berryvilleiml.com/taxonomy/

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," arXiv Preprint, https://arxiv.org/abs/1412.6572, 2014.

[12] Xinyun Chen et al., "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," *arXiv [cs.CR]* (December 15, 2017), arXiv, http://arxiv.org/abs/1712.05526.

[13] Yisroel Mirsky et al., "CT-GAN: Malicious Tampering of 3D Medical Imagery Using Deep Learning," *arXiv [cs.CR]* (January 11, 2019), arXiv, http://arxiv.org/abs/1901.03597.

[14] Nicholas Carlini et al., "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," *arXiv [cs.LG]* (February 22, 2018), arXiv, http://arxiv.org/abs/1802.08232.

[15] Charles Perrow, *Normal Accidents: Living with High Risk Technologies* (Princeton University Press, 2011); Scott Douglas Sagan, *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons* (Princeton University Press, 1995).

[16] "WannaCry: Ransomware Attacks Show Strong Links to Lazarus Group" (Symantec, May 22, 2017), https://www.symantec.com/connect/blogs/wannacry-ransomware-attacks-show-strong-links-lazarus-group.

[17] Robert Lee, Michael Assante, and Tim Conway, "SANS ICS Defense Use Case (DUC) Dec 30 2014: ICS CP/PE Case Study Paper-German Steel Mill Cyber Attack," 2014, https://ics.sans.org/media/ICS-CPPE-case-Study-2-German-Steelworks_Facility.pdf.

[18] James Bamford, "The Most Wanted Man in the World" (Wired, August 22, 2014), https://www.wired.com/2014/08/edward-snowden/.

[19] Kim Zetter, *Countdown to Zero Day* (New York: Crown, 2014).

[20] "What's the Worst That Could Happen?" (Government Communications Headquarters, 2016), https://www.documentcloud.org/documents/2699620-What-Is-the-Worst-That-Can-Happen-March-2010.html.

[21] Robert S. Mueller III, "Report On The Investigation Into Russian Interference In The 2016 Presidential Election" (Department of Justice, March 2019), https://www.justice.gov/storage/report.pdf.

[22] Radford et al., "Language Models Are Unsupervised Multitask Learners."

[23] Sarah Kreps and Miles McCain, "Not Your Father's Bots," *Foreign Affairs*, April 16, 2020, https://www.foreignaffairs.com/articles/2019-08-02/not-your-fathers-bots.

[24] Daniel M. Ziegler et al., "Fine-Tuning Language Models from Human Preferences," *arXiv [cs.CL]* (September 18, 2019), arXiv, http://arxiv.org/abs/1909.08593.

[25] For these quotes and discussion of this idea, see Buchanan, *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations*, chap. 2.

[26] Robert Jervis, "Cooperation Under the Security Dilemma," *World Politics* 30, no. 2 (1978): 167–214.

[27] Ben Buchanan, "The Life Cycles of Cyber Threats," *Survival* 58, no. 1 (2016), http://www.iiss.org/en/publications/survival/sections/2016-5e13/survival--global-politics-and-strategy-february-march-2016-44d5/58-1-03-buchanan-7bfc.

[28] Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare* (Profile Books, 2020).

[29] Thucydides, *History of the Peloponnesian War*, trans. Richard Crawley (New York: Dover, 2012).

[30] Jervis, "Cooperation Under the Security Dilemma"; Charles Glaser, "The Security Dilemma Revisited," *World Politics* 50, no. 1 (1997): 171–201; Buchanan, *The Cybersecurity Dilemma: Hacking, Trust and Fear Between Nations*, chap. 2.